SEMPEDIA : Sémantisation à partir des documents semi-structurés - Enrichissement de DBPédia

Nathalie Aussenac-Gilles, Cécile Fabre, Adel Ghamnia, Mouna Kamel, Cassia Trojahn dos Santos

▶ **To cite this version:**

Nathalie Aussenac-Gilles, Cécile Fabre, Adel Ghamnia, Mouna Kamel, Cassia Trojahn dos Santos. SEMPEDIA : Sémantisation à partir des documents semi-structurés - Enrichissement de DBPédia. [Contrat] Université de Toulouse-le-Mirail. 2020. hal-02960440

HAL Id: hal-02960440
https://hal.science/hal-02960440v1

Submitted on 7 Oct 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

SEMPEDIA :

Sémantisation à partir des documents semi-structurés - Enrichissement de DBPédia

Rapport sur les travaux de thèse d'Adel Ghamnia

Sous la direction de Cécile Fabre et Nathalie Aussenac-Gilles

Rapport de fin de contrat de la région Midi-Pyrénées
Convention 620402C526
Université Toulouse 2 Jean Jaurès - juillet 2019

Auteurs :
Nathalie Aussenac-Gilles (IRIT, CNRS)
Cécile Fabre (CLLE, Université Toulouse - Jean Jaurès)
Adel Ghamnia (IRIT, CLLE, UT2J)
Mouna Kamel (IRIT, université de Perpignan)
Cassia Trojahn (IRIT, Université Toulouse - Jean Jaurès)

Résumé

Cette thèse s'inscrit dans le cadre d'un projet interdisciplinaire appelé SemPedia qui a fait l'objet d'un partenariat entre les équipes Melodi et ERSS de l'Institut de Recherche en Informatique de Toulouse (IRIT) et du laboratoire Cognition, Langue, Langage, Ergonomie (CLLE). L'objectif est de mettre en commun des compétences en informatique, linguistique et traitement automatique des langues pour le développement d'outils d'extraction de connaissances à partir de textes visant l'enrichissement du Web des données en français. Le Web des données permet la publication de données structurées pour constituer un réseau de connaissances à l'échelle du Web. Or, les contenus relatifs au français restent très insuffisants par rapport à leurs équivalents en anglais. Ainsi, la ressource DBPedia en français est 20 000 fois plus pauvre que la version anglaise de DBPedia. Ce rapport rend compte des méthodes mises au point dans le cadre du projet SemPedia pour la construction automatique de ressources sémantiques à grande échelle pour le français. Elles reposent sur le développement de plusieurs extracteurs de connaissances à partir du corpus Wikipedia, constitué de textes de nature encyclopédique particulièrement riches en informations. Cette ressource textuelle a été généralement exploitée par des techniques analysant seulement la partie la plus fortement structurée de la base (infoboxes, catégories, etc.), délaissant la majorité de l'information textuelle disponible. Le projet SemPedia a pour objectif d'évaluer les apports de méthodes qui visent à tirer parti de toutes les informations textuelles disponibles dans le corpus Wikipedia en combinant des approches variées d'extraction des connaissances, basées à la fois sur des règles linguistiques (patrons morpho-syntaxiques) et sur des processus d'apprentissage (supervision distante).

Le projet s'est focalisé sur l'extraction de connaissances de nature taxonomique, qui constituent l'ossature principale des ressources sémantiques. Elles organisent les concepts sous forme de hiérarchies, selon une relation dite de spécialisation en intelligence artificielle, appelée hyperonymie en linguistique. L'extraction de relations à partir des textes a motivé de nombreux travaux, conduisant à un foisonnement de méthodes qui s'explique à la fois par la disparité des situations d'acquisition et par les évolutions technologiques. La thèse implémente plusieurs techniques complémentaires, afin d'évaluer les conditions optimales de leur utilisation, en fonction du type de ressources textuelles disponibles (plus ou moins normalisées et structurées) et de la nature de l'information textuelle mobilisée (contenu textuel standard, titres, énumérations, etc.).

Ce rapport présente la problématique de la thèse, fait un bilan des travaux antérieurs sur cette question, et détaille les différents volets de la contribution en faisant varier plusieurs paramètres : la nature des textes (textes spécifiques ou ensemble du corpus Wikipedia), les caractéristiques de l'information textuelle (texte brut ou structuré), et les types de techniques mobilisés (approche par règles ou par apprentissage).

Les chapitres relatifs à la contribution présentent et intègrent 4 des 5 publications réalisées dans le cadre du projet.

Mots clés : Web des données, extraction de connaissances, bases de connaissances, traitement automatique des langues, relations sémantiques

Table des matières

1 Introduction

1.1 Le projet SemPedia

Le web des données (*Linked Open Data*) a pour vocation de publier à grande échelle des données structurées sur le web et de les relier, permettant ainsi aux machines d'interpréter et d'exploiter ces informations. Le web des données forme ainsi un réseau de connaissances, qui est devenu un vecteur de développement et de diffusion incontournable. Par exemple, avec le *Knowledge Graph* [Singhal, 2012-05-16] puis le *Knowledge Vault* [Dong et al., 2014], le moteur Google utilise ce type de données pour désambiguïser les requêtes et afficher une synthèse des connaissances associées aux entités répondant à la requête. Dans le domaine médical, ce type de ressource facilite l'exploitation automatique de résumés scientifiques.

1.1.1 DBpédia en français

Dans ce contexte, il apparaît crucial d'enrichir le web des données pour la langue française, car ce type de ressource fait encore défaut aujourd'hui. La mise à disposition de la ressource *DBpédia en français*[1] en 2013 va dans ce sens. Mais cette ressource n'est actuellement alimentée que par des informations extraites de données fortement structurées (infobox, catégories, liens, etc.) provenant des pages de Wikipedia en français, sous-exploitant le potentiel d'information présent dans les contenus textuels associés aux pages. En effet, les pages de Wikipedia, de nature encyclopédique, contiennent des développements en langue naturelle qui sont riches en connaissances mais qui ne se retrouvent ni dans les infobox ni dans les catégories Wikipedia ni dans les liens entre pages. Ce texte exprime des concepts (de domaine ou généraux), des relations et des règles qui les associent et leur donnent du sens. De plus, les pages Wikipedia ont la particularité d'être semi-structurées, c'est-à-dire qu'elles combinent mise en forme (paragraphes, titres, listes énumératives, etc.) et texte linéaire. Les attributs de mise en forme du texte, qui organisent et contribuent largement à sa sémantique, sont rarement pris en compte par les méthodes classiques du traitement automatique des langues.

Afin de répondre au besoin d'exploiter le texte des pages Wikipédia pour enrichir DBPedia en français, deux équipes de recherche toulousaines, l'équipe ERSS du laboratoire CLLE (UMR5263) et l'équipe MELODI de l'IRIT (UMR 5505), ont mis en place le projet SemPedia, dont ce rapport présente les travaux et les résultats.

1.1.2 Objectifs de SemPédia

Le projet SemPedia a deux objectifs.

Le premier objectif est de proposer un ensemble de méthodes et d'outils qui combinent analyse de la langue et mise en forme du texte, pour extraire les relations sémantiques entre termes, et plus spécifiquement les relations d'hyperonymie (généricité/spécificité) à partir d'éléments de structure courants tels que les titres, listes, tableaux, paragraphes. La focalisation sur ces relations hiérarchiques se justifie par leur rôle essentiel dans la modélisation de connaissances. Elles constituent l'ossature principale des ressources sémantiques, ossature sur laquelle s'appuient généralement les processus de raisonnement.

Le second objectif est d'enrichir ensuite le web des données pour la langue française, en particulier DBPedia en français, en utilisant ces extracteurs sur les pages Wikipedia en

1. http ://fr.dbpedia.org/

français pour former des triplets reliant des ressources (classes ou entités ou valeurs) par des prédicats étiquetés (ou relations).

Le projet SemPedia soulève donc des problèmes de traitement massif de données textuelles du web, et participe à la mise au point de ressources sémantiques pour l'accès aux connaissances du web. De nature interdisciplinaire, il sollicite l'expertise conjointe de chercheurs en linguistique et en ingénierie des connaissances des deux laboratoires. Il vient à la fois renforcer un sujet de recherche commun, l'extraction de relations sémantiques, et inaugurer de nouvelles études en développant la prise en compte de la structure textuelle, la distribution des mots dans de grands corpus et les liens avec les ressources sémantiques du web. Ce projet s'inscrit aussi dans la continuité des travaux menés sur la sémantique textuelle depuis les années 1990 à Toulouse en sciences cognitives (réseau PRESCOT), dans lesquels les deux laboratoires ont fortement collaboré.

1.2 Focalisation sur l'extraction de relations

Dans le cadre du projet SemPedia, la thèse d'Adel Ghamnia contribue à la problématique de l'extraction de relations sémantiques à partir du texte des pages WikiPedia. Plus particulièrement, nous nous focalisons sur les relations binaires, celles reliant deux termes en corpus, et qui seront ensuite représentées par un triplet reliant deux entités ou deux classes ou une entité et une classe ou encore une classe et une valeur. Formellement, une relation binaire est un triplet $r = (e_1, e_2)$ où r est le prédicat (que nous appellerons désormais nom ou type de la relation), et e_i des termes (en général des syntagmes nominaux) désignant les éléments en relation. Par exemple, la phrase (exemple 1.2) permet d'identifier plusieurs relations binaires comme `est-un(blues, genre musical)`. `est-un(blues, genre vocal)`, `est-un(blues, genre instrumental)`, `est-dérivé(blues, chants de travail des populations afro-américaines)`, `pays-origine(blues, États-Unis)`, ... D'emblée, cette phrase montre la difficulté de la tâche d'extraction de relations, tant pour les repérer dans les textes (par exemple, les mots "genre" et "vocal" ne sont pas contigus dans la phrase) que pour les représenter (il faut faire une sorte de déduction et un choix pour décider de représenter les États-Unis comme "pays-d'origine" du blues alors que la phrase dit que ce sont les populations qui l'ont créé qui étaient aux États-Unis ; la relation pourrait s'appeler "a-pour-origine" ou encore "origine").

(exemple 1.2) *Le blues est un genre musical, vocal et instrumental dérivé des chants de travail des populations afro-américaines subissant la ségrégation raciale aux États-Unis.*
`est-un(blues, genre musical).` `est-un(blues, genre vocal),` `est-un(blues, genre instrumental),` `est-dérivé(blues, chants de travail des populations afro-américaines)` , `pays-origine(blues, États-Unis)`, ...

Pratiquement, l'extraction de relations a fait l'objet de recherches au sein de plusieurs communautés scientifiques : en **traitement automatique des langues** (TAL) comme sous-tâche de l'extraction d'information, en terminologie et en **ingénierie des connaissances** (IC), lorsque des textes sont utilisés comme une des sources pour construire des terminologies formelles et des bases de connaissances.

1.2.1 Extraction de relations et extraction d'information

En TAL, elle est considérée comme une sous-tâche du processus d'extraction d'information (EI) qui cherche à découvrir un nombre fini et réduit de relations entre les mentions linguistiques de deux (ou plus) entités de types prédéfinis (et toujours en nombre limité, à partir de textes en langage naturel [Culotta et al., 2006]. Les premiers travaux de ce type remontent aux années 1990 et ont porté sur l'analyse de dépêches financières Grishman [1997]. On s'intéresse par exemple à des relations de localisation entre des personnes et des lieux, des personnes et des dates (de naissance ou de mort) ; des événements et des dates et/ou des lieux, entre des personnes et des entreprises, etc. Ces informations sont ensuite stockées dans des bases de données, dont chaque table est associée à un type de relation, les types de données correspondant aux classes recherchées et chaque ligne étant une instance de relation. Au sujet des travaux pionniers menés sur le français, le livre de Thierry

Poibeau présente une synthèse remarquable [Poibeau, 2003]. Le domaine évolue pour prendre en compte des problèmes plus complexes car il bénéficie désormais des algorithmes d'apprentissage automatique. Afin de dégager le plus grand nombre de triplets des textes, l'extraction d'information "ouverte" ne fait aucun a priori sur les classes recherchées ou sur les relations possibles entre ces classes [Etzioni et al., 2008].

L'extraction d'information est un domaine de recherche en pleine effervescence, d'une part à cause des avancées très prometteuses permises par la représentation des mots sous forme de vecteurs de plongement en corpus (word embeddings) et d'autre part grâce aux algorithmes d'apprentissage automatique, en particulier à base de réseaux de neurones.

Parmi les états de l'art offrant une vue synthétique des avancées du domaine, citons celui de Bach and Badaskar [2007], suivi par l'article de Sharma et al. [2016] sur l'extraction de relations binaires dans le domaine médical. Plus récemment, Pawar et al. [2017] a fait un panorama des méthodes d'apprentissage supervisé et semi-supervisé alors que Smirnova and Cudré-Mauroux [2018], Niklaus et al. [2018] et Kumar [2017] utilisent respectivement la supervision distante pour faire de l'extraction d'information ouverte et l'apprentissage profond pour l'extraction de relations. Enfin, un bon état de l'art sur l'extraction d'information en général et l'extraction de relations en particulier dans le cadre du web a été rédigé par Martinez-Rodriguez et al. [2018].

1.2.2 Extraction de relations en Terminologie et en Ingénierie des connaissances

En Terminologie, l'organisation des terminologies dans des bases de données puis dans des bases de connaissances a d'abord donné lieu à des travaux sur leur représentation et sur les schémas de données associés, mais très vite aussi sur la manière de les construire plus ou moins automatiquement à partir de textes. Les travaux d'Ingrid Meyer dans les années 1990 sont pionniers en la matière [Skuce and Meyer, 1990] [Meyer et al., 1992] avec la notion de bases de connaissances terminologique. Faisant l'hypothèse que les corpus contribuent à "révéler" les termes et leur signification dans un domaine particulier, I. Meyer a proposé de s'appuyer sur les textes pour identifier des termes en usage et de tenir compte de leurs contextes pour rendre compte de leur sens et de leurs relations. L'extraction de termes et l'extraction de relations sont alors devenues deux tâches clés du processus de construction de terminologie, alors qu'il était jusque là plutôt manuel et lié à l'expertise de spécialistes du domaine. Les premiers logiciels réalisant ces extractions s'appuient sur des patrons lexico-syntaxiques, en particulier ceux de M. Hearts pour la construction de la taxonomie des termes [Barrière, 2004], que l'on retrouve opérationnels dans le logiciel TerminoWeb [Barrière and A., 2006]. Un état de l'art sur l'extraction de relations en terminologie est paru en 2008 [Auger and Barrière, 2008].

En IC, l'extraction de relations à partir de textes s'intéresse avant tout aux relations binaires car les langages de représentation des connaissances Extraire des relations sémantiques de textes est un moyen de collecter des connaissances et de construire ces modèles, qui intervient dans deux tâches liées au processus de construction de modèles de connaissances à partir de textes [Buitelaar et al., 2005] [Buitelaar and Cimiano, 2008] [Lehmann and Volker, 2014] :

- La construction d'ontologie ou de schéma de modèle. Il s'agit ici d'identifier les relations potentielles entre classes, qui peuvent être des relations hiérarchiques (des relations de spécialisation entre classes, de partie-tout entre classes) ou non (relations de causalité, d'origine, datation, localisation, etc.). En général, les classes sont extraites en étudiant les termes spécifiques au domaine utilisés en corpus, et les types de relations sont découverts au fur et à mesure de l'extraction des relations elles-mêmes, ce qui conduit à un processus itératif. Une attention particulière est portée à la relation d'hyperonymie. Lorsqu'elle relie deux classes, elle est représentée à l'aide de la relation `rdfs:subClassOf` entre classes. Cette relation organise les concepts en une taxonomie qui sert d'ossature à toute ontologie. L'apprentissage de cette taxonomie à partir de texte fait l'objet de recherches spécifiques sous le nom de "Taxonomy learning" [Wang et al., 2017].
- la construction de bases de connaissances ou peuplement d'ontologies. Dans ce cas, on peut faire l'hypothèse que les classes à renseigner sont connues (et le problème se rapproche alors de celui de l'extraction d'information) ; ou pas connues (on cherche alors de manière "ouverte" toutes les relations possibles entre entités (et dans ce cas, on se rapproche de travaux récents en *open information extraction*).

Parmi les travaux pionniers, citons deux systèmes basés sur l'utilisation de patrons lexico-syntaxiques, Prométhée d' E. Morin [Morin and Jacquemin, 2004], qui permet d'apprendre de nouveaux patrons spécifiques à un corpus et Caméléon de P. Séguéla [Séguéla and Aussenac-Gilles, 1999], qui accompagne le processus de la mise au point (manuelle) de patrons spécifiques à l'enrichissement d'un modèle conceptuel en passant par l'adaptation et la validation des patrons

aux corpus et aux connaissances à décrire [Séguéla, 2001]. Cette approche a été reprise et complétée dans une deuxième version de Caméléon [Jacques and Aussenac-Gilles, 2006] avec des patrons appliqués aux résultats d'un corpus étiqueté syntaxiquement.

Cette problématique fait aujourd'hui l'objet de nombreuses recherches et d'avancées régulières dans l'état de l'art. Parmi les résultats marquants des années 2000 à 2010, citons Text2Onto [2] [Cimiano and Völker, 2005] intégré dans la plate-forme NEON de construction d'ontologies et de bases de connaissances, et l'utilisation de la plate-forme GAte [3] [Cunningham et al., 2002] [Cunningham and Bontcheva, 2013] pour définir des chaînes d'extraction de relations et d'entités. En effet, que ce soit pour la construction d'ontologies ou pour l'enrichissement de bases de connaissances, les relations sémantiques jouent un rôle fondamental pour organiser les concepts d'un domaine et pour former un graphe de connaissances par la mise en relation des concepts et entités. Les travaux récents s'appuient sur l'apprentissage automatique et les réseaux de neurones, et l'on assiste à une multiplication des résultats et des approches pour la partie extraction proprement dite, alors que moins de recherches portent sur la génération et l'organisation en graphe de triplets RDF.

Nous mentionnerons les travaux les plus avancés dans l'état de l'art du chapitre 2. Ce qu'il ressort des premiers états de l'art [Bach and Badaskar, 2007], c'est que les techniques d'apprentissage les plus performantes sont des techniques supervisées, qui requièrent de disposer d'exemples positifs et négatifs. Ces exemples doivent être annotés comme comportant une relation, et laquelle, ou aucune relation, permettant ainsi de produire une classification en 2 ou n classes suivant le nombre de relations recherchées. Or l'annotation de fragments de texte est un travail long et fastidieux, conduisant parfois à des désaccords entre experts. Pour éviter cette phase, des alternatives consistent à (i) produire des exemples automatiquement à partir de patrons, c'est la supervision partielle comme dans l'approche Snorkel [4] [Ré, 2018] [Hancock et al., 2018] ou (ii) à exploiter des ressources pour annoter (bases de connaissances contenant déjà des relations validées), c'est la supervision distante. Elle a été appliquée à l'extraction de relations dès 2005 environ [Mintz et al., 2009] [Min et al., 2013].

Cette idée est mise en oeuvre dans les travaux les plus avancés qui utilisent aussi des plongements vectoriels de mots (word embeddings) et des réseaux de neurones. Une des ressources utilisées est par exemple FreeBase dans [Xu and Barbosa, 2019]. On trouvera un panorama des systèmes et approches les plus récents sur les deux sites suivants :

— "https ://github.com/roomylee/awesome-relation-extraction"
— "https ://nlpprogress.com/english/relationship-extraction.html"

Bien sûr, ces travaux ont été testés essentiellement sur des jeux de données (textes et triplets en relation) en anglais et nécessitent d'importants volumes de textes pour entraîner les modèles.

1.3 Objectifs de la thèse

La thèse ne vise pas la définition de nouvelles approches plus performantes, mais plutôt la validation de l'hypothèse que chacune des approches de l'état de l'art ne permet pas de trouver les mêmes occurrences de relation dans les textes, et donc que l'utilisation cumulée de plusieurs approches assure un meilleur rappel (permet de trouver plus de relations valides) que chacune des approches prise séparément. Nous voulons montrer expérimentalement qu'il est préférable de combiner plusieurs techniques complémentaires afin de mieux exploiter différentes spécificités des textes analysés.

Cette hypothèse a été motivée par plusieurs travaux de l'état de l'art. Tout d'abord, l'idée d'exploiter une multiplicité de techniques reprend l'approche mise en œuvre pour la construction de YAGO, avec la notion *d'extracteur* [Suchanek et al., 2007] spécialisé dans l'extraction de certains types de connaissances, chacun selon une approche spécifique. Ensuite, elle prolonge la thèse de Roger Leitzke Granada, qui a évalué finement la complémentarité de 4 techniques d'extraction de relations d'hyperonymie [Leitzke Granada, 2015] et montré que chacune permettait de cibler un ensemble différent de relations entre entités. Enfin, nous intégrons le travail de Jean-Philippe Fauconnier et Mouna Kamel [Fauconnier and Kamel, 2015], qui ont montré que l'exploitation de la structure des textes permet d'extraire des relations qui échappent à l'exploitation du seul contenu textuel linéaire.

Formulée autrement, notre hypothèse se décline ainsi :

— le fait de multiplier différentes techniques appliquées aux mêmes documents permet de repérer avec chacune différentes expressions linguistiques de relations sémantiques. On peut

2. http ://neon-toolkit.org/wiki/1.x/Text2Onto.html
3. https ://gate.ac.uk/
4. https ://www.snorkel.org/

ainsi extraire un plus grand nombre de relations. Nous nous proposons donc d'implémenter plusieurs techniques complémentaires pour extraire des relations d'hyperonymie.
— le fait d'adapter une technique aux spécificités de formulation ou de mise en forme de certaines pages rend l'extraction de relations plus efficace. Nous avons donc identifié des sous-ensembles de Wikipedia présentant des particularités qui puissent rendre leur exploitation ciblée plus efficace, et nous avons défini des techniques adaptées.
— le fait d'exploiter à la fois la structure d'un texte et son contenu textuel permet de traiter plus efficacement le contenu des pages que si l'on exploite seulement le contenu non structuré. Pour cela, nous avons cherché à implémenter des approches exploitant des éléments structurés des textes (comme les listes) en plus du contenu textuel *stricto sensu*.

1.4 Organisation du document

La suite du rapport est organisée comme suit :
— Le **chapitre 2** est consacré à un état de l'art présentant les enjeux de l'enrichissement des bases de connaissances et les méthodes déployées dans les travaux antérieurs.
— Le **chapitre 3** présente les résultats produits pendant une première phase de la thèse, visant la mise en œuvre de méthodes à base de patrons, testées sur plusieurs échantillons de données. L'objectif est d'évaluer les performances d'une approche simple à base de règles, combinant des patrons génériques et des patrons ajustés aux caractéristiques des documents considérés.
— Le **chapitre 4** présente une deuxième méthode recourant à l'apprentissage supervisé, et plus précisément au principe de supervision distante.
— Le **chapitre 5** généralise les résultats obtenus en étendant la taille du corpus analysé et en combinant les deux approches précédentes.
— Le **chapitre 6** présente un prolongement de l'approche définie dans le cadre de la thèse, appliquée cette fois aux structures énumératives.
— Le **chapitre 7** présente les conclusions et les perspectives de travail à venir.

Les chapitres ayant donné lieu à publication sont constitués d'une synthèse préalable à la présentation de l'article publié. Voici le détail des articles publiés dans le cadre du projet SemPedia :
— Ghamnia, A. (2016). Extraction de relations d'hyperonymie à partir de Wikipédia. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2016, volume 3 : RECITAL*, Paris, France, 40-51.
— Ghamnia, A., Kamel, M., Trojahn, C., Fabre, C., Aussenac-Gilles, N. (2017). Extraction de relations : combiner les techniques pour s'adapter à la diversité du texte. In *Actes des 28es Journées francophones d'Ingénierie des Connaissances IC 2017*,Caen, France. 86-97.
— Kamel, A., Trojahn, C., Ghamnia, A., Aussenac-Gilles, N., Fabre, C. (2017). Extracting hypernym relations from Wikipedia disambiguation pages : comparing symbolic and machine learning approaches. In *Proceedings of the International Conference on Computational Semantics (IWCS)*, Montpellier, France.
— Kamel, A., Trojahn, C., Ghamnia, A., Aussenac-Gilles, N., Fabre, C. (2017). A Distant Learning Approach for Extracting Hypernym Relations from Wikipedia Disambiguation Pages. In *Proceedings of the 21st International Conference on Knowledge-Based and Intelligent Information Engineering Systems (KES 2017)*, Marseille, France. 1764-1773.
— Kamel, A., Trojahn, C. (2018). Towards Enriching DBpedia from Vertical Enumerative Structures Using a Distant Learning Approach. In *Proceedings of the International Conference on Knowledge Engineering and Knowledge Management (EKAW 2018)*, Nov 2018, Nancy, France. pp.179-194.

2 Etat de l'art

Sommaire

Ce chapitre présente l'état de l'art des travaux relatifs à notre étude. Dans une première partie, nous faisons le point sur les notions fondamentales d'ontologie et de base de connaissances, et présentons quelques grands bases de connaissances partagées et largement utilisées. Nous insistons particulièrement sur DBPédia en français puisque nous voulons enrichir cette base de connaissances. Dans une deuxième partie, nous faisons le point sur les travaux récents en matière d'extraction de relations à partir de textes. Nous nous intéressons particulièrement aux approches à base de patrons et celles utilisant l'apprentissage automatique semi-supervisé.

2.1 Contexte : Ontologies et Bases de connaissances

L'un des objectifs de cette thèse est d'enrichir les bases de connaissances, spécialement DBpedia pour le français. Les bases de connaissances sont des modèles formels permettant de stocker des données sémantiques ou linguistiques. Elles ont été proposées dans l'objectif de leur appliquer des mécanismes automatiques de raisonnement au sein d'un Système à base de connaissance, qui est principalement constitué d'une base de connaissances (BC) et d'un moteur d'inférences. Historiquement, les premières bases de connaissances datent des années 60, et utilisent une représentation des connaissances à base de logique, puis de règles de production [Minsky, 1974].

Depuis la proposition du langage RDF [1] par le W3C (organisme de standardisation du web), les BCs sont plus souvent représentées par un graphe dont les nœuds sont les notions manipulées (concepts, entités ou valeurs), et les arêtes représentent les relations qui relient ces notions. Par exemple, dans la relation :

`[voiture] 'a pour date de construction' [date]`

`[voiture]` et `[date]` sont des concepts qui seront des noeuds du graphe, et 'a pour date de construction' une relation qui va étiqueter l'arrête du graphe reliant ces deux noeuds. `[voiture]` et `[date]` sont des concepts car ils renvoient à des ensembles d'objets du monde réel possédant les mêmes propriétés.

Les noeuds d'un graphe peuvent aussi représenter des instances, c'est-à-dire des objets spécifiques du monde réel qui peuvent appartenir à une de ces classes. Par exemple, `[voiture]` et `[date]` pourraient avoir respectivement pour instances `[maClio]` et `[2000]`. Une partie du graphe peut aussi exprimer la relation `[maClio]` 'a pour date de construction' et `[2000]`.

Un des travaux fondateurs de cette approche est la recherche de J. Sowa sur les Graphes Conceptuels [Sowa, 1984]. Il introduit cette représentation de connaissances pour rendre compte formellement de la sémantique de phrases en langage naturel. D'autres travaux de référence ont utilisé la logique, comme dans les réseaux sémantiques [Shapiro, 1971] ou encore les frames [Minsky, 1974]. Depuis les années 90, les recherches en matière de représentation des connaissances se

1. https ://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/

sont focalisées sur la description des entités d'un domaine nécessaires pour une application. Les modèles ainsi constitués ont été appelés *ontologies* (informatiques) au sens où elles représentent les catégories sémantiques essentielles qui seront manipulées par les raisonnements.

Nous présentons dans cette section (i) une définition des ontologies et leurs modèles et (ii) un état de l'art des bases de connaissances.

2.1.1 Ontologies et Formalismes

Les ontologies sont aujourd'hui utilisées dans plusieurs domaines d'application tels que le Web Sémantique, le traitement du langage naturel et la fouille d'informations. Plusieurs définitions présentent une ontologie comme une conceptualisation qui explicite la sémantique d'un domaine défini par une communauté de pratique ou par les usagers d'applications informatiques. Nous retenons une définition qui distingue les ontologies de domaine des autres modèles informatiques tels que les modèles conceptuels et les modèles de connaissance.

Définitions

Une des définitions les plus citées est celle de T. Gruber Gruber [1993], selon laquelle

> une ontologie est une conceptualisation :
> — Formelle : exprimée dans un langage syntaxique et sémantique formalisé (RDF, RDF Schema, OWL) qui permet de réaliser des raisonnements et déductions pour vérifier la consistance ou inférer de nouveaux faits.
> — Consensuelle : acceptée par tous les membres du domaine.
> — Référençable : tout concept décrit par l'ontologie possède un identifiant, qui permet de le référencer à partir de n'importe quel contexte.

Plus théoriquement, N. Guarino Guarino [1998] introduit la notion d'ontologie ainsi :

Dans son usage le plus courant en IA, une ontologie réfère à un artefact d'ingénierie, consitué d'un vocabulaire spécifique utilisé pour décrire une certaine réalité, plus un ensemble d'affirmations qui donnent du sens aux mots de ce vocabulaire. Cet ensemble d'affirmation prend habituellement la forme d'une théorie en logique du premier ordre, où les mots du vocabulaire apparaissent comme des prédicats unaires ou binaires, appelés respectivement concepts et relations. [2]

Il en propose alors une définition plus précise à l'aide des notions de *conceptualisation* (comme approximation du réel) et surtout d'*engagement ontologique* (qui est une interprétation intentionnelle qui rend explicite la manière dont un langage L peut être approché par une conceptualisation C) :

> Une ontologie est une *théorie logique* qui rend compte du sens que l'on veut donner à un *vocabulaire formel*, c'est-à-dire *son engagement ontologique* pour une conceptualisation particulière du monde. Les *modèles intentionnels* du langage logique utilisant ce vocabulaire sont contraints par son engagement ontologique. Une ontologie reflète donc indirectement cet engagement (et la conceptualisation sous-jascente) en représentant de manière approchée les modèles souhaités.

Pratiquement, cette définition insiste plus que celle de Gruber sur la théorie logique associée au vocabulaire : pour définir une ontologie, il ne suffit pas de lister des concepts et des relations possibles entre ces concepts, il faut aussi expliciter sous forme de formules logiques le sens qu'on leur donne, c'est-à-dire les inférences que l'on peut faire à partir de ce vocabulaire. les règles et les axiomes d'une ontologie, ainsi que la sémantique des relations, sont donc fondamentaux.

2. In its most prevalent use in AI, an ontology refers to an engineering artifact, constituted by a specific vocabulary used to describe a certain reality, plus a set of explicit assumptions regarding the intended meaning of the vocabulary words. This set of assumptions has usually the form of a first-order logical theory, where vocabulary words appear as unary or binary predicate names, respectively called concepts and relations.

Principaux composants des ontologies

Les ontologies sont basées sur cinq principaux composants : les classes, les propriétés, les types de valeurs, les axiomes et les instances.

Classes : Appelée aussi Concept, une classe est une description abstraite et commune de plusieurs objets semblables. Elle possède, d'une part, un identifiant et une description textuelle ayant pour but de la rattacher à une connaissance préexistante de l'utilisateur et, d'autre part, des relations formelles avec d'autres composants de l'ontologie.

Propriétés : Une propriété est un composant qui permet de caractériser et de distinguer les instances d'une classe. Une propriété possède un identifiant, une définition comportant une partie textuelle et une partie formelle, et un type qui peut être défini dans un domaine de valeurs simples ou de classes.

Types de valeurs : Les types de valeurs définissent les ensembles de valeurs sur lesquels les propriétés prennent leurs valeurs. Ces ensembles incluent toujours (1) les types simples (Integer, Boolean, String,....), (2) les classes (ce type de propriétés représente des relations et associations) et (3) des collections.

Axiomes : Les axiomes sont des prédicats qui s'appliquent sur les classes ou les instances des classes d'une ontologie, permettant de restreindre les interprétations d'une ontologie et d'inférer de nouveaux faits. Les principaux axiomes exprimés par les langages de définition des ontologies sont les axiomes de typage, subsomption et de cardinalité.

Instances : Appelés aussi objets ou individus, les instances représentent des éléments spécifiques d'une classe. Une instance appartient à une (ou plusieurs) classes et ses propriétés peuvent prendre des valeurs.

Ontologie Conceptuelle

Une **Ontologie Conceptuelle** (OC) est une ontologie dont le but est la représentation des catégories de concepts et des propriétés des concepts présents dans un domaine, contrairement aux Ontologies Linguistiques qui ont comme objectif de représenter la signification des termes utilisés dans un langage naturel donné.

Concepts des ontologies conceptuelles

Les concepts définis dans une ontologie conceptuelle ont deux types [Gruber, 1993] :

1. **Concepts primitifs** : ces concepts ne peuvent pas être définis par une définition axiomatique complète, ces concepts représentent une base sur laquelle d'autres concepts peuvent être définis. Par exemple, on peut décider de représenter `Humain, Male, Femelle` comme des concepts primitifs.
2. **Concepts définis** : ces concepts sont définis par des formules logiques à l'aide des concepts primitifs. Par exemple, à partir des concepts primitifs `Humain, Masculin et Féminin`, on peut définir les concepts suivants : `Femme = Humain ET Féminin`, `Homme = Humain ET Masculin`. Ces concepts sont alors des concepts définis.

Types d'ontologie conceptuelle

Selon [Fankam et al., 2008], on peut distinguer deux types d'ontologie conceptuelle :

1. Ontologies Conceptuelles Canoniques (OCC) : ontologies dont les définitions ne contiennent aucune redondance, chaque concept de ces ontologies ne peut être décrit que d'une seule façon. Par conséquent, une OCC ne contient que les concepts primitifs.
2. Ontologies Conceptuelles Non Canoniques (OCNC) : ontologies qui contiennent des concepts définis en plus des concepts primitifs. Ces ontologies sont en particulier utilisées pour définir des mappings entre différentes ontologies.

Formalismes de définition des ontologies

Plusieurs formalismes ont été proposés pour concevoir les ontologies, nous présentons dans cette section les principaux utilisés dans le cadre du web sémantique : RDF, RDF-Schema, OWL. Ce sont les trois standards du W3C conçus pour permettre d'associer aux documents du Web des données sémantiques exploitables par une machine. Ces standards sont largement décrits dans plusieurs articles et documents. Une introduction très pédagogique en est faite dans le RDF-PRIMER `https://www.w3.org/TR/rdf11-primer/` . Les documents produits par le W3C pour spécifier ces langages sont des recommandations disponibles en ligne :

— la sémantique de RDF est en ligne ici : `https://www.w3.org/TR/rdf11-mt/`
— la recommandation définissant RDFS est ici : `https://www.w3.org/TR/rdf-schema/`
— celle de OWL :`https://www.w3.org/TR/owl2-overview/`

1. Pour représenter des connaissances, RDF utilise des triplets pour relier des ressources du web au sein d'un graphe. Chaque noeud du graphe est une ressource, identifiée par un identifiant unique du web, une URI. et chaque arc du graphe est un prédicat, également identifié par un identifiant URI. les unités de connaissances sont des triplets <Sujet,Prédicat,Objet> où :
   — Le sujet est une ressource identifiée par une URI (Uniform Resource Identifier) .
   — Le prédicat est une propriété qui caractérise cette ressource, également identifiée par une URI.
   — L'objet désigne une 2e ressource, qui peut être soit une URI, soit une valeur, soit un noeud vide.

   Par défaut, RDF propose quelques prédicats réservés :
   — rdf :type qui permet de spécifier qu'une ressource est le type d'une autre ressource ;
   — rdf :label qui permet d'associer une chaîne de caractères à une ressource ou à un prédicat ;
   — rdf :property qui permet de spécifier un prédicat associé à une ressource.
2. RDF a été étendu par RDF-Schema (RDFS) qui reprend les prédicats et principes de RDF et fournit des prédicats supplémentaires afin de distinguer plusieurs types de ressources : les classes (ou concepts) et les entités (ou instances). On admet que ces prédicats permettent de représenter des ontologies simples, essentiellement les hiérarchies de concepts grâce aux prédicats rdfs :subClassOf et rdfs :subProperty :
   — rdfs :class permet de définir une ressource comme étant une classe ; ainsi, on peut distinguer des entités, indivudus, et des classes, ensembles d'individus. Le prédicat rdf :type est utilisé pour dire qu'une ressource est un individu d'une classe.
   — rdfs :subClassOf permet de spécifier une organisation hiérarchique des classes : on indique qu'une classe est incluse dans une autre.
   — rdfs :subProperty qui permet de spécifier une organisation hiérarchique des propriétés.
   — rdfs :domain qui permet d'associer une propriété à une classe.
   — rdfs :range qui permet de spécifier le co-domaine de valeur d'une propriété
3. OWL est conçu afin d'étendre RDF et RDF-Schema de manière à décrire des ontologies avec une sémantique plus riche et sur le Web. Le principal objectif d'OWL est d'assurer les déductions et les inférences. Pour cela, ce langage offre la possibilité de définir des concepts primitifs ou définis, des axiomes et des contraintes sur les propriétés. OWL [3] s'appuie toujours sur une représentation par triplets (la syntaxe de RDF) et peut organiser les classes en hiérarchies (comme RDFS).

   En plus des primitive de RDFS, OWL permet d'exprimer que deux classes sont disjointes (`owl:disjointClass`), que deux classes sont équivalentes (`owl:equivalentClass`) ou que deux entités sont identiques (`owl:sameAs`) ; il permet de distinguer de propriétés pointant vers une classe (`owl:ObjectProperty`) ou une valeur (`owl:DataTypeProperty`) ; il fournit des constructeurs de classes complexes, des restrictions sur les propriétés etc.

   La définition de OWL se heurte à a difficulté de tout langage de représentation des connaissances : trouver un compromis entre expressivité et capacité de raisonnement. En effet, plus un langage de représentation permet des représentations riches et fines (par exemple, possibilité d'exprimer des disjonctions, des négations, des quantificateurs existentiels ou universels), plus il est difficile de prouver qu'un ensemble de connaissances exprimées avec ce langage est cohérent. En cas d'inconsistance, il n'est pas possible de raisonner de manière fiable à partir d'un ensemble de faits, ce qui rend presque inutilisable la base de connaissances. De ce fait, dès sa première version proposée à partir de 2004, ce formalisme a donné lieu à trois variantes qui offrent plus ou moins de capacités d'expression : OWL Lite (langage minimal), OWL DL (équivalent aux logiques de description) et OWL Full (propose tous les opérateurs mais non vérifiable).

   Une deuxième version, OWL-2, est désormais la recommandation du W3C depuis 2009, actualisée en 2012 (les documents de référence sont présentés ici `https://www.w3.org/TR/2012/REC-owl2-overview-20121211/`. Les ontologies en OWL2 peuvent prendre du sens selon deux sémantiques : celle de OWL2-DL (restriction de la spécification de OWL2 de manière à faciliter les raisonnements automatiques, définie par une sémantique appelée sémantique directe qui correspond à une représentation en logique de description) et OWL2-Full (spécification complète ed OWL2 qui étend la sémantique de RDFS et considère une ontologie en OWL2 comme un graphe RDF).

3. https ://www.w3.org/2001/sw/wiki/OWL

La spécification de OWL2 permet de définir de nouvelles variantes (ou profils) de OWL2. Cependant, 3 autres profils sont définis et souvent utilisés par défaut (cf `https://www.w3.org/TR/2012/REC-owl2-primer-20121211/`) :

— OWL2-EL est proche de OWL2-DL. Il est adapté à la représentation de très grandes bases de connaissances contenant beaucoup de classes, de relations rdfs :subClassOf et d'instance. On peut exprimer en OWL-EL la négation, la disjonction, des axiomes et des contraintes sur les propriétés, mais pas la notion de propriété inverse.
— OWL2-RL est plus complet tout en étant calculable. Il peut être exprimé à l'aide d'un langage de règles (Rule Language).
— OXL2-QL rend compte de toutes les informations que l'on trouve habituellement en UML et dans les langages exprimant les schémas de BD. Il a une expressivité équivalent à celle de SQL.

2.1.2 Quelques bases de connaissances

Plusieurs types d'applications nécessitent de disposer de modèles de bases de connaissances : pour caractériser le contenu de textes, raisonner ou rechercher des informations par exemple. Dans l'optique du web sémantique, ces ressources sont un moyen de disposer d'un "vocabulaire" formel pour annoter des pages web, c'est-à-dire associer de classe à de chaîne de caractères. Nous présentons ici plusieurs base de connaissances générale qui présentent un intérêt pour notre projet : DbPedia car c'est la base de connaissances dont nous voulons enrichir la version française en traitant automatiquement le contenu de pages Wikipedia, Yago car cette base intègre DBPedia et d'autres connaissances, et enfin BabelNet, qui reprend également les différentes versions de DBPedia dans différentes langues et des connaissances directement extraites de pages web pour former une base de connaissances multilingue.

DBpedia

Wikipedia est une encyclopédie universelle, en ligne et libre d'accès qui traite tous les domaines de connaissances. Elle est organisée en pages, chaque page contenant des informations sur une connaissance particulière et pouvant être éditées collaborativement par toute personne qui le souhaite via Internet et en respectant la charte Wikipedia. Chaque page et rattachée à une ou plusieurs catégories qui caractérisent son contenu. Ces catégories sont organisées dans une hiérarchie des plus générales aux plus spécifiques. La nature de ces catégories est définie par les utilisateurs, en fonction des connaissances à publier.

Chaque page répond à un plan type qui est propre à sa catégorie. De plus toutes les pages ont en commun (1) de commencer par un paragraphe qui tient lieu de définition de la notion présentée sur cette page ; (2) de comporter une synthèse des éléments essentiels à la notion présentée, dans un encadré appelé InfoBox.

DBpedia [4] est un projet qui a pour objectif de créer une base de connaissances qui représente les connaissances décrites dans Wikipedia, plus particulièrement le contenu des InfoBox. Le projet produit une structure formelle, une base de connaissances (BC), également appelée DBPedia, produite automatiquement à partir des éléments structurés de Wikipedia : la hiérarchie de catégorie et les contenus des infoboxes en particulier. Cette BC est définie par une ontologie qui définir les classes des entités décrite dans les infoboxes, ainsi que les entités utilisées dans ces descriptions. L'ontologie DBpedia contient 320 concepts décrits par 1600 propriétés, extraits des éléments structurés des pages Wikipedia via différents outils dédiés appelés « extracteurs ». Les pages Wikipedia contiennent des textes et des éléments de structure qui définissent la connaissance décrite.

Les extracteurs DBpedia sont de quatre types :

— *Extracteur Infobox* : ce type d'extracteur réalise le mapping entre l'Infobox ciblée et les termes de l'ontologie DBpedia. Ainsi, chaque entrée (ligne) de l'infobox est associée à une propriété de la la classe de l'ontologie décrivant la notion présentée dans l'infobox. Il existe un extracteur de ce type pour chacune des catégories Wikipedia car leur infoboxe suit un modèle (template) différent pour chaque catégorie.
— *Extracteur de lignes Infobox* : ce type d'extracteur relie directement chaque entrée (ligne) de l'infobox aux concepts ou instances de l'ontologie DBpedia.
— *Extracteurs ciblés* : ces extracteurs permettent d'extraire des connaissances à partir d'une structure ciblée de page Wikipedia qui respecte la charte Wikipedia, comme par exemple : les tables, les données géographiques, lles catégories, etc.

4. https ://wiki.dbpedia.org/

FIGURE 2.1 – Une page Wikipédia

— *NLP extracteurs* : ces extracteurs, basés sur l'apprentissage automatique, permettent d'extraire des entités nommées : comme par exemple les noms de personnes, de lieux, etc. Ces extracteurs utilisent une approche statistique basée sur un algorithme de clasification SVM.

Il existe plusieurs versions de DBPedia, dont une pour le français : DBPediaFR [5]. C'est cette version que nous visons à enrichir dans cette thèse. En effet, comme nous venons de le présenter, les seules sources utilisées par les extracteurs sont les infoboxes. DBpedia intègre aussi la hiérarchie de catégories de Wikipedia. Mais le texte des pages, pourtant riche en connaissances, n'est pas exploité. Notre projet SemPedia vise justement à exploiter ces textes pour y représenter des connaissances complémentaires à celles déjà présentes dans les infoboxes. Nous espérons ainsi enrichir la base de connaissances.

Yago

Yago [6] est une BC qui contient aussi le résultat de l'extraction de connaissances à partir des pages Wikipedia, couplé au contenu de la base de données lexicales WordNet [7], et plus récemment GeoNames [8]. Yago est donc différente de DBpedia. De plus, la manière d'exploiter Wikipedia est un peu différente, et cela grâce à un grand nombre d'extracteur spécialisés. Comme DBpedia, Yago s'appuie sur la structure des Catégories de Wikipedia. Les Catégories de Wikipedia regroupent tous les articles qui appartiennent à cette catégorie. Par exemple, la page `https://fr.wikipedia.org/wiki/ZinÃĭdine_Zidane` est dans la catégorie Footballeurs Français. Dans cette catégorie, on trouve des instances (dont 'Zidane') du concept 'Footballeur'. La problématique traitée par Yago est définie par l'hypothèse : « In an ontology, concepts have to be arranged in a taxonomy to be of use". Or, dans Wikipedia, les catégories ne sont pas directement des Concepts, et l'arbre des catégories n'a pas rigoureusement une structure de hiérarchie de classes : sous-classes. L'hypothèse d'ontologie n'est donc pas vérifiée. La solution proposée pour déduire que « Footballeurs Français » est un « Footballeur », est d'utiliser la ressource linguistique Wordnet, où sont formalisées les relations lexico-conceptuelles entre les mots d'une langue. Ainsi, Yago présente une approche qui combine à la fois les pages WIkipedia et la ressource Wordnet, et donne une précision de 97%.

Pour construire Yago, le choix a été fait représenter par une instance le sujet de chaque page Wikipedia. Pour définir la classe à associer à chaque instance candidate, un module appelé *Wikipedia*

5. http ://fr.dbpedia.org/
6. https ://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/
7. fabian2007yago
8. yago2016

Category System exploite la structure de Catégorie de Wikipedia. Pour identifier les catégories conceptuelles, le nom de la catégorie candidate est analysé grâce au parseur *Noun Group Parser* qui le découpe en (i) pré-modifieur, (ii) entête et (ii) post-modifieur, par exemple : *Naturalized citizens of the United States* devient : « *Naturalized* », « *citizens* » et « *United* States ». Selon cette procédure, si l'entête extraite est un nom pluriel, il est possible qu'il soit un concept. Cela est validé si cet entête appartient à un synset de Wordnet. Pour définir la nature de la relation entre une instance et un concept, Yago traite aussi chaque catégorie de Wikipedia, par exemple si le nom de la catégorie se termine par le suffixe « _births », la relation est «BornInYear».

BabelNet

BabelNet [Navigli and Ponzetto, 2012] est une ressource multilingue qui peut être vue comme une grande encyclopédie numérique, une ressource lexicale ou une base de connaissances représentées par un graphe. BabelNet a été construit en combinant l'utilisation des pages de Wikipedia et de la ressource lexicale Wordnet. A partir de Wordnet, tous les mots (*senses*) de WordNet sont considérés comme des concepts et tous les liens entre ces mots sont considérés comme des relations sémantiques. A partir de Wikipedia, tous les intitulés des pages donnent lieu à la création de concepts et tous les liens hypertextes entre ces pages sont considérés comme des relations entre les concepts. Pour produire une seule base de connaissances, BabelNet propose d'utiliser un mapping entre chaque page Wikipedia et son terme équivalent dans Wordnet dans toutes les langues possibles. Le problème soulevé par cette approche est celui de la désambiguïsation. En effet, le risque de relier un terme t présent dans Wordnet à une mauvaise page Wikipedia est élevé. Pour cela, l'algorithme de création de Babelnet exploite tous les termes appartenant au synset du terme t dans WordNet en considérant les différentes relations qui les relient : synonymie, hyperonymie, etc. Une fois que les instances et concepts sont définis, les autres structures Wikipedia (infobox, hiérarchie des catégories, etc) sont exploitées pour en extraire d'autres relations (lieu de naissance, etc.). Les résultats de l'évaluation de BabelNet font état d'un apport conséquent pour chaque langue (allemand, français, espagnol, italien et catalan). BabelNet peut être consultée en ligne `https://babelnet.org/`, interrogée via une API ou exploitée dans une application.

2.1.3 Construction d'ontologies à partir de textes

Les textes sont utilisés comme sources de connaissances pour construire des bases de connaissances depuis la fin des années 90, avec la disponibilité massive de corpus numériques d'une part, et les avancées des logiciels de traitement automatique des langues d'autre part [Maedche and Staab, 2000] [Buitelaar et al., 2005] [Bourigault et al., 2004], mais aussi par la disponibilité de ressources linguistiques et sémantiques générales (comme WordNet, BabelNet et Yago) qui améliorent les performances du traitement du langage.

Des logiciels spécialisés ont été développés pour des tâches propres à la construction d'ontologies : extracteurs de concepts (i.e. Syntex [Bourigault, 2002] ou SystemQuirk [Ahmad and Holmes-Higgin, 1995]) et extracteurs de relations sémantiques (i.e. Prométhée [Morin, 1999] et Caméléon [Séguéla, 2001]). Ces travaux visent toujours des domaines de spécialité. Leur succès a conduit à de nouveaux développements pour intégrer différentes techniques et logiciels au sein de plates-formes. C'est ainsi qu'un module TAL, OntoLT, a été proposé pour l'éditeur Protégé [Buitelaar et al., 2003], un autre pour WebODE [Arpirez et al., 2003], et que l'atelier KAON d'édition d'ontologie a intégré Text-to-Onto [Cimiano and Völker, 2005] pour extraire des éléments d'ontologie des textes. A partir de 2003, on a assisté à une véritable explosion de l'exploitation des textes, avec deux tendances fortes : l'utilisation du web comme corpus et la recherche systématique de l'automatisation, en particulier par des techniques d'apprentissage automatique et d'extraction d'information [Buitelaar et al., 2005]. On a parlé alors d'ontology learning [Maedche and Staab, 2000] [Buitelaar et al., 2006]. Dans ce cadre, l'apprentissage associé à des techniques d'extraction d'information et d'annotation sert à retrouver dans des textes des traces linguistiques de la présence de concepts ou d'instances de concepts, ce qui permet de compléter l'ontologie par des classes, des propriétés et des axiomes pour la construction de l'ontologie ou des instances et des relations entre instances pour le peuplement de l'ontologie.

Nous présentons l'approche dans sa généralité, pour nous focaliser ensuite sur la tâche qui nous concerne plus précisément, à savoir l'extraction de relations sémantiques. De même que l'on distingue l'ontologie (ensemble des classes, des axiomes et donc des relations entre classes du modèle à construire) de la base de connaissances (ensemble des entités du domaine et de leur relations décrits avec les classes et les propriétés définis dans l'ontologie), l'état de l'art différencie les processus de construction et enrichissement d'une ontologie d'une part, et celui de peuplement de l'ontologie (ou peuplement de la base de connaissances) d'autre part.

Des textes aux ontologies et BC

Construire une base de connaissances à partir de textes fait appel à des logiciels et de principes méthodologies de deux types :

- des principes et logiciels d'extraction d'information à partir de textes :
  - extraction des termes significatifs du domaine, qui désigneront ensuite soit des classes, soit des instances de ces classes : on peut ici utiliser soit des logiciels de TAL de base (analyseurs pour découper le textes en unités linguistiques, analyseurs syntaxiques pour identifier les catégories grammaticales etc), accessibles dans des plateformes comme GATE [Cunningham et al., 2002] soit des extracteurs de termes spécialisés comme TerminoWeb [Barrière and A., 2006] ou TermoStat (droin), des extracteurs d'entités nommées (NER) [Kazama and Torisawa, 2007] ; une autre approche consiste à retrouver dans les textes des termes tirés de bases de connaissances du domaine étudié ;
  - extraction de relations sémantiques entre termes, dans le but de définir ensuite des relations sémantiques entre noeuds du graphe ; nous détaillons ces travaux par la suite. Pour la construction d'ontologies, on cherche des relations entre classes, donc plutôt entre des syntagmes nominaux basés sur des noms communs, alors que le peuplement consiste à chercher des relations entre instances. Les recherches en extraction d'informations sont très réutilisées ici pour le peuplement : étant données des classes de l'ontologie et les relations que l'on s'attend à trouver entre elles, on cherche ces relations précisément entre des instances de ces classes [Nédellec et al., 2009], [Buitelaar et al., 2005], [Buitelaar and Cimiano, 2008].
  - plus récemment, des étapes de filtrage, nettoyage et correction des données extraites permet soit d'entraîner des algorithmes d'apprentissage pour poursuivre l'extraction, soit de retirer des erreurs et du bruit dans les résultats.
- des principes, langages et logiciels de modélisation des connaissances extraites : la méthode la plus utilisée et citée est Methontologie associée à la plateforme Neon ; la méthode Terminae [Aussenac-Gilles et al., 2008], elle, s'appuie sur une analyse de la terminologie.

L'accent est souvent mis sur les logiciels d'extriment au détriment de la phase de modélisation, avec une hypothèse forte et pourtant naïve que l'on pourrait directement représenter au sein d'un graphe des éléments de connaissances en fonction de leur structure linguistique, et que ceux-ci soient pertinents pour l'application dans laquelle ce graphe va être utilisé. Les travaux menés sur Caméléon [Séguéla, 2001] puis sur la méthode Terminae [Aussenac-Gilles et al., 2008] soulignent le processus d'interprétation et de sélection, souvent implicite, associé à la modélisation, ainsi que les limites des textes comme source de connaissances d'un domaine, ou encore le faible rappel et le bruit des logiciels d'extraction, qui conduisent à des modèles très imparfaits.

Bien sûr, depuis l'arrivée des algorithmes à base de réseaux de neurones et l'explosion du volume de textes disponibles, ces modèles sont de plus en plus précis et la qualité des résultats en est meilleure. Mais la manière d'intégrer l'information extraite dans des modèles conduit encore à des redondances ou à des incohérences comme on les constate dans DBpedia. Pour améliorer l'approche, d'un processus linéaire tel qu'il était défini dans les années 2010, on est passé aujourd'hui à un processus itératif, faisant appel à plusieurs techniques qui se complètement mutuellement, et à l'intégration d'étapes de nettoyage des résultats appris, de correction ou de filtrage, pour ré-entraîner les modèles et parvenir à de meilleurs résultats. C'est ce que propose par exemple la plate-forme SNORKL [Ré, 2018].

Peuplement de bases de connaissances

Le peuplement d'une BC sert à ajouter à une BC de nouvelles entités et relations entre entités en respectant la structure de son ontologie (le cas de DBPedia, Yago et BabelNet).

Tout comme la construction de l'ontologie à partir de textes, l'enrichissement d'une base de connaissances qui contient déjà des classes et des définitions de propriétés exploite les textes généralement en deux étapes : extraire d'abord des données conceptuelles à partir de traces linguistiques, et les consolider par rapport à l'ontologie de la BC, à savoir les associer à la bonne classe de l'ontologie, les placer au bon endroit du graphe et cela en cohérence avec les informations déjà présentes dans la base de connaissances.

La première étape consiste généralement à exploiter des documents textuels pour en extraire des instances, dont les marques linguistiques sont parfois des syntagmes nominaux, et le plus particulièrement des entités nommées, et des relations entre instances. Nous présenterons les approches d'extraction de relations dans la partie suivante. La deuxième étape de consolidation des résultats de l'extraction sert à vérifier la non-redondance de la connaissance et à préserver la cohérence et la qualité de la BC. Pour cela, un ensemble de règles de consolidations doit être défini selon la structure de l'ontologie. Ces règles visent à vérifier qu'une instance de même classe et

même libellé n'existe pas déjà dans la BC, que des héritages multiples ne sont pas contradictoires, que les valeurs multiples d'un même attribut ne sont pas incompatibles ou encore que la base de comporte pas de cycle, etc.. Dans le cas où l'information n'est pas valide, il reste la possibilité de la corriger manuellement ou la négliger.

Un état de l'art de 2011 [Petasis et al., 2011] fait ressortir trois approches qui traitent cette problématique :

— Artequakt system [Kim et al., 2002] appliquent des patrons définis à la main pour extraire des connaissances et les introduire dans la BC. L'élimination des informations redondantes se fait à la fin du processus.
— SOBA system [Buitelaar et al., 2006], suivent le même principe mais en validant l'information au moment de l'annotation pour éviter les redondances.
— BOEMIE [Castano et al., 2008] utilisent l'apprentissage automatique pour extraire des informations et des techniques d'alignement pour les ajouter à une ontologie en évitant d'introduire des informations redondantes.

Discussion : WIkipedia comme corpus

WIkipedia offre d'énormes opportunités comme corpus pour l'extraction d'informations de toutes sortes : concepts, relations, faits, affirmations, axiomes ou règles. L'article très complet de Medelyan et al. [2009] montre comment, dès 2000, ces éléments de connaissances peuvent servir à l'analyse du langage naturel, la recherche d'information, l'extraction d'information, mais aussi la construction et le peuplement et la construction d'ontologies. Wikipedia est largement utilisé pour construire des ressources génériques. Historiquement, la première initiative de grande ampleur a été DBPedia [Morsey et al., 2012], puis Yago [Suchanek et al., 2007], [Rebele et al., 2016] et ensuite BabelNet [Navigli and Ponzetto, 2012]

Ces travaux d'extraction de relations d'hyperonymie à partir de Wikipédia ciblent particulièrement les parties structurées : la définition [Kazama and Torisawa, 2007] [Navigli and Velardi, 2010] ; les menus [Sumida and Torisawa, 2008], qui offrent un moyen d'accéder à la hiérarchie des concepts ; les catégories [Navigli and Ponzetto, 2012] et es infoboxes [Morsey et al., 2012].

En revanche, le texte qui forme le corps des pages est utilisé moins efficacement. La partie appelée "définition", qui est le premier paragraphe de chaque page, est presque systématiquement traité car elle permet de situer le concept décrit sur la page par rapport à des concepts plus généraux ou plus précis. Mais le texte rédigé est peu exploité, en particulier dans la construction des BCs actuelles et en cure plus pour lespages en français. Cette section doit introduire le besoin d'exploiter les documents textes présenté dans la section suivante.

2.2 Extraction de relations à partir de textes

L'objectif de cette thèse est d'extraire des relations à partir de documents semi-structurés pour enrichir des Bases de Connaissances. Nous présentons dans ce chapitre les différentes approches d'extraction de relations à partir de textes écrits en langage naturel. Ces approches ont toutes pour objectif de modéliser les relations sémantiques exprimées par le langage naturel, mais elles se différencient par la méthode utilisée pour l'extraction de relations. En premier temps, nous présentons l'approche d'extraction dite linguistique, principalement fondée sur la définition de patrons lexico-syntaxiques caractérisés par différents marqueurs présents dans le texte. En second lieu, nous présentons l'approche dite statistique, fondée sur l'apprentissage automatique.

2.2.1 La relation d'hyperonymie

La définition aristotélicienne de l'hyperonymie est la suivante :
«L'hyperonymie est une **fonction** qui, à partir d'un **terme** *t*, retourne un ou plusieurs **termes** plus généraux.»

L'hyperonymie est donc une relation qui sert à hiérarchiser des objets (pouvant être des termes ou des concepts). Dans le domaine de l'ingénierie des connaissances, l'hyperonymie représente l'ossature de toute ontologie. Elle est souvent assimilée à la relation *is-a*, par exemple par [Brachman and Schmolze, 1989], qui la définissent comme suit :

«si *X* est une classe d'objets, et *X'* une sous-classe de *X*, alors *is-a*(*X'*, *X*) est vrai»

L'argument le plus général dans une relation d'hyperonymie est appelé *hyperonyme*, tandis que l'argument le plus spécifique est appelé *hyponyme*. Pour le reste de cette thèse, nous retenons la formulation suivante de l'hyperonymie :

$$hyperonymie(x, y) = \begin{cases} 1 & si\ hyperonyme(x) = y \\ 0 & sinon \end{cases}$$

Sur le plan conceptuel, il existe trois niveaux dans la hiérarchie de la relation d'hyperonymie, le niveau médian correspondant à l'usage le plus courant [Kleiber and Tamba, 1990] (Table 2.1) :

Niveaux	Exemples de domaines	
niveau générique	plante	fruit
niveau de base	arbre	pomme
niveau spécifique	chêne	reinette

TABLE 2.1 – Les niveaux de la relation hiérarchique d'hyperonymie

Nous considérons l'ensemble de ces niveaux dans la thèse.

Sur le plan linguistique, la relation d'hyperonymie peut être exprimée de façon variée, dans des structures combinant des indices lexicaux et morphosyntaxiques. Pour le français, ces marqueurs de l'hyperonymie ont été recensés par [Borillo, 1996], dans la continuité des travaux de Hearst que nous mentionnons juste après [Hearst, 1992] : la structure prédicative attributive (Nx est un Ny : *le bacille est une bactérie*) est la plus évidente, car elle est présente dans les définitions des dictionnaires. Néanmoins elle ne permet pas de marquer de façon non ambigüe la relation. Elle est complétée par un ensemble de structures coordonnées avec un marqueur de spécification (ex : Nx ou tout autre Ny : *un ulcère ou toute autre maladie de l'estomac*), des structures appositives (Nx, Ny : *les batraciens, animaux à sang froid*), des structures d'inclusion (parmi Ny, Nx : *parmi les roches volcaniques, la domite*), des structures d'énumération (Ny tels que Nx1, Nx2, Nx3...), et des structures nominales de la forme Ny de Nx (*un sentiment de colère*) - on parle dans ce cas également d'inclusion lexicale. On voit donc que le lexique, les structures syntaxiques mais aussi la ponctuation participent à la formulation de la relation d'hyperonymie. Des structures plus complexes impliquant la dimension discursive (organisation inter-phrastique) ont été mises en évidence, comme le recours à l'anaphore associative [Condamines and Jacques, 2006].

2.2.2 Approche linguistique

Par approche linguistique, nous entendons ici toute approche des données textuelles qui repose sur la prise en compte de connaissances linguistiques dans le traitement. Une approche fondatrice est celle introduite par Hearst [1992]. Cette approche est basée sur des patrons lexico-syntaxiques. Un patron lexico-syntaxique correspond à une caractérisation abstraite de toutes les réalisations langagières associées à la relation, et peut être composé d'éléments lexicaux, grammaticaux ou sémantiques [Auger and Barrière, 2008].

Un exemple de patron proposé par Hearst et qui permet d'extraire une relation d'hyperonymie entre X et Y est :

Y such as X ((, X) (, and/or) X).*

Ce patron permet, par exemple, d'identifier la relation d'hyperonymie entre *animal* et *cat*, entre *animal* et *mouse*, et entre *animal* et *dog* dans la phrase *animals such as cats, mice and dogs are classified as...* (après lemmatisation).

Les travaux de Hearst ont été étendus par d'autres travaux que nous présentons un peu plus loin. L'appellation de *patron lexico-syntaxique* n'a pas été partagée. Notamment, Séguéla and Aussenac-Gilles [1999] parlent de *formules linguistiques*, [Meyer, 2001] de *Knowledge Patterns* (patrons de connaissances) et Arnold and Rahm [2015] de *Semantic Patterns* (patrons sémantiques).

D'après [Auger and Barrière, 2008], il existe deux approches pour construire des patrons lexico-syntaxiques :

- L'approche *onomasiologique* : consiste à identifier les marqueurs qui expriment une relation donnée en analysant les phrases qui révèlent cette relation pour proposer le patrons le plus adéquat.
- L'approche *sémasiologique* : consiste à identifier les relations exprimées par un ensemble de marqueurs fixes.

Les travaux de [Hearst, 1992] ont été étendus pour identifier des relations d'hyperonymie entre termes [Morin and Jacquemin, 2004], [Séguéla, 2001] et les ont adaptés pour le français. [Berland and Charniak, 1999] ont aussi proposé des patrons pour une autre relation qui est la méronymie.

Ces travaux ont montré qu'il est possible d'adapter les patrons pour une relation donnée ou une langue donnée. Les résultats obtenus montrent généralement une bonne précision mais un faible rappel. [Séguéla, 2001] démontre que la bonne précision des patrons est due au fait que l'on privilégie plutôt la qualité des patrons à leur quantité, ce qui explique aussi leur faible rappel.

En contrepartie des ces avantages, les patrons présentent un inconvénient important qui est le coût très élevé de leur construction. La définition manuelle est généralement à refaire pour chaque nouveau domaine et nécessite une expertise humaine de ce domaine Jacques and Aussenac-Gilles [2006].

Plusieurs travaux ont essayé de remédier à cet inconvénient par la proposition d'outils assistants à la création de patrons. [Morin, 1999] a proposé le système PROMÉTHÉE qui utilise des patrons pour extraire des relations sémantiques basés sur l'analyse distributionnelle de chaque domaine visé. [Séguéla, 2001] a proposé l'outil CAMELEON qui permet d'acquérir des connaissances d'un domaine et d'adapter leur base de patrons aux marqueurs extraits de ce domaine. En complément des outils eux-mêmes, les ressources de patrons se sont progressivement enrichies. Ainsi, [Lefeuvre, 2017] a proposé une base de patrons pour différentes relations sémantiques en français et en anglais, dont l'hyperonymie. Nous avons réutilisé ces différentes bases de patrons dans nos travaux, et les présentons dans le chapitre suivant.

Assurant la réutilisabilité des ressources et facilitant la construction de nouveaux patrons, ces outils ont permis de diminuer le coût lié à la construction de patrons. Néanmoins, la question de la couverture des marqueurs exprimant chaque relation cible reste ouverte : les constructions visées ne rendent compte que d'une part très restreinte des contextes dans lesquels est susceptible d'apparaitre une paire de termes en relation d'hyperonymie. Ce constat explique l'évolution des travaux réalisés dans le domaine au profit de l'approche dite statistique, basée sur l'apprentissage automatique, que l'on présente dans la section suivante.

2.2.3 Approche statistique

D'un point de vue général, l'apprentissage automatique sert à concevoir et implémenter des méthodes qui permettent à une machine d'adapter ses analyses et ses comportements en se fondant sur l'analyse empirique de données. L'objectif de l'apprentissage automatique est de produire des règles, sans intervention humaine, en se basant sur des exemples annotés [Mitchell, 1997]. Il est impliqué dans de nombreux secteurs tels que : la recherche d'informations, la classification de données, l'extraction d'informations, etc.

A partir de cette définition, l'approche statistique pour l'extraction de relations à partir de textes vise à :

— apprendre les marqueurs à partir de régularités présentes dans un texte,
— généraliser ces marqueurs sur de nouvelles données textuelles,
— extraire des relations à partir de ces nouvelles données.

Il existe trois types d'apprentissage automatique dans l'approche statistique : l'apprentissage supervisé, l'apprentissage semi-supervisé et l'apprentissage non supervisé.

Apprentissage supervisé

L'apprentissage supervisé est une méthode qui utilise un algorithme de classification déjà entraîné sur un ensemble d'exemples annotés, généralement à la main. Les exemples d'entraînement, appelés *training set* sont généralement décrits par différents traits. Pour l'analyse du langage naturel les traits peuvent être des traits de bas niveau (configuration de n-grammes) ou des traits faisant appel à des informations linguistiques plus ou moins sophistiquées : informations morpho-syntaxiques (catégories grammaticales), syntaxiques (liens de dépendance), voire sémantiques. Pour évaluer l'algorithme de classification entraîné sur des exemples et utilisant ces traits, on utilise un ensemble de test, appelé *test set*.

Dans la littérature, on retrouve deux types de méthodes d'apprentissage supervisé pour extraire des relations à partir du langage naturel :

Les méthodes à base de traits : Comme leur nom l'indique, ces méthodes se caractérisent par l'exploitation de traits. Tout d'abord, pour toute donnée issue de l'ensemble d'entraînement on associe un vecteur de traits. Ce vecteur de traits est composé d'un ensemble de traits pouvant être de nature syntaxique, morphologique, sémantique, etc. Ensuite, le classifieur qui aura pour tâche de classer les nouvelles données devra se baser sur ce vecteur de traits pour déterminer la classe de ces nouvelles données. La définition d'un bon ensemble de traits peut être réalisée à l'aide d'un algorithme de sélection de traits [Chen et al., 2005]. Les travaux qui utilisent ces méthodes se différencient généralement par la relation cible et par le classifieur utilisé : (Kabhlatla, 2004) utilisent un modèle log-linéaire basé sur la régression

logistique multinomiale, [Zhao and Grishman, 2005] et [Zhou et al., 2005] utilisent un SVM. Tandis que [Deschacht and Moens, 2006], [Anke, 2013] utilisent un CRF.

Les méthodes à base de noyaux : Au lieu de générer des vecteurs de traits, on utilise une fonction noyau qui calcule la similarité entre les arbres syntaxiques des exemples d'entraînement. Leur principal avantage se présente dans le fait qu'il n'est pas nécessaire de définir "manuellement" un ensemble de traits. La fonction noyau permet de définir la classe de toute nouvelle donnée en analysant simplement à partir de son arbre syntaxique et ceux des exemples d'apprentissage. Parmi les travaux qui utilisent ces méthodes on retrouve : (Zalenko et al., 2003) qui proposent une fonction noyau sur des arbres de constituants syntaxiques, [Culotta et al., 2006] et [Bunescu and Mooney, 2005] qui utilisent des arbres de dépendances syntaxiques, et (Zhang et al., 2006) et (Choi et al., 2009) qui proposent des fonctions noyaux composites.

Une des limites de l'apprentissage supervisé est la construction de l'ensemble des exemples annotés. La définition de cet ensemble d'apprentissage (qui est généralement faite à la main) s'avère très coûteuse. L'apprentissage semi-supervisé et non supervisé visent à réduire ce coût.

Apprentissage semi-supervisé

Deux méthodes sont proposées dans cette approche pour réduire le coût de la construction de l'ensemble d'exemples d'entraînement : l'amorçage et la supervision distante.

L'amorçage (*bootstrapping*) L'amorçage consiste à créer automatiquement un ensemble d'entraînement à partir d'un petit sous-ensemble d'exemples annotés à la main (appelé *amorces*). Dans le contexte de l'extraction de relations, les amorces peuvent être les arguments de la relation cible, ou un patron correspondant à la relation cible. Ces amorces sont projetées sur le corpus, et de nouvelles relations (resp. de nouveaux arguments), sont identifiées. Le processus peut être réitéré.

(Hearst, 1998) a proposé cette méthode pour apprendre des patrons lexico-syntaxiques. Ce travail a été étendu par (Brin, 1998), qui définit cette méthode pour extraire des relations entre *auteur* et *livre* pour construire l'outil *DIPRE*. L'inconvénient de cette méthode est le bruit très élevé qui pourrait être généré. En effet, les erreurs d'extraction pouvant apparaitre dans l'itération *n*, apparaissent forcément dans l'itération *n+1*. D'autre part, cette méthode génère aussi des dérives sémantiques (*semantic drifts*) : lorsque le sens des arguments n'est pas pris en compte dans l'itération *i*, il génère des erreurs dans l'itération *n+1*. Des travaux ont essayé de proposer des solutions à ce problème (Kozareva et al., 2008 ; Hovy et al., 2009 ; Kozareva et Hovy, 2010) en introduisant des labels d'arguments dans les patrons, par exemple le patron *type such as seed and ** devient *Presidents such as Ford and X**. D'autres travaux ont été proposé pour régler ce type de problème en intégrant une mesure de similarité entre contextes syntaxiques des patrons (Zhang et al., 2014) et (Lambrou-Latreille, 2015).

L'amorçage a pu montrer une réduction de coût d'annotation des exemples d'entraînement. Mais il nécessite quand même la définition des amorces, qui s'avère parfois difficile en l'absence d'expertise humaine liée aux types de corpus. D'autre part, la faible précision due au bruit qui peut s'accumuler au fur et à mesure des itérations demeure un problème. Pour remédier à cela, la supervision distante a été introduite.

La supervision distante La supervision distante est une approche d'apprentissage qui repose sur l'hypothèse suivante : « Si on sait que deux entités sont reliées par une relation r, n'importe quelle phrase qui contient ces deux entités pourrait exprimer la relation r» [1] [Mintz et al., 2009]. Cette approche combine une base de connaissances externe (qui fournit des propositions de relations entre entités) et un corpus pour créer son ensemble d'apprentissage : les paires d'entités en relation dans la base de connaissances sont "projetées" sur les phrases du corpus. Toutes les phrases contenant deux entités reliées par la relation r dans la base de connaissances sont considérées comme des exemples positifs de la relation r. Cela présente deux avantages :

— plus besoin d'amorces (contrairement au bootstrapping)
— plus besoin de corpus annoté (contrairement à l'apprentissage supervisé)

Les différents travaux relevant de cette approche se différencient essentiellement par la construction des exemples d'apprentissage, par les entités considérées, et par les algorithmes d'apprentissage mis en œuvre.

La supervision distante est une méthode largement répandue aujourd'hui. Les relations trouvées entre deux entités textuelles de la base de connaissances sont associées aux phrases où ces entités apparaissent. L'ensemble d'exemples positifs d'apprentissage est donc composé de ces phrases. Les entités de la base de connaissances dont aucune relation ne les relient sont associées aux

phrases où ces deux entités apparaissent, mais pour construire l'ensemble d'exemples négatifs d'apprentissage. L'hypothèse de base [1] ne pouvant être toujours vérifiée, Riedel et al. [2010] proposent l'introduction d'un modèle graphique pour modéliser cette incertitude. En effet, cette hypothèse augmente le bruit ramené par les phrases qui n'expriment pas la relation R mais contiennent les deux entités textuelles X et Y.

L'hypothèse de base de leur travail devient alors :

«Un exemple $e = (X, Y, R, s)$ tel que $s = (s_1, s_2, .., s_n)$, où s_i est une phrase qui contient X et Y, est classé vrai-positif si $\exists s_i$ telque s_i contient un terme qui exprime la relation R.»

Par exemple, les deux arguments de la paire (*Roger MacNamee*,*Elevation Partners*) (X et Y) sont reliés par la relation *founded* : en effet, le terme *founded* exprime la relation *is_founder*. Les deux phrases :

(1) «*Elevation Partners*, the $1.9 billion private equity group that was founded by private the *Roger MacNamee*..»

et

(2) «*Roger MacNamee*, a mananging director at *Elevation Partners*..»

contiennent les deux arguments X et Y mais la phrase (2) n'exprime pas la même relation R car elle ne contient pas le terme "*founded*".

Les recherches actuelles dans la supervision distante essaient d'affiner cette hypothèse (Linge et al., 2013) et de mieux sélectionner les phrases d'entraînement (Xu et al., 2013).

Une autre approche qui s'affranchit de l'annotation manuelle des exemples d'entraînement est basée sur l'apprentissage non supervisé. Son avantage majeur est de ne pas cibler une relation particulière à extraire.

Apprentissage non supervisé

Cette approche repose essentiellement sur le principe de *clustering* : ces méthodes cherchent à décomposer un ensemble d'individus en plusieurs sous ensembles les plus homogènes possibles. On cherche alors à maximiser la similarité intra-classe et à minimiser la similarité inter-classes. Plusieurs méthodes existent : elles se différencient essentiellement par la mesure de distance (qui dépend du type de données). Par exemple, dans le cadre de l'extraction de relations, les vecteurs de traits associés aux relations identifiées en corpus peuvent faire l'objet de la classification.

En se basant sur ce principe, (Hasegawa et al. 2004) ont présenté une méthode de *clustering* dont le critère de regroupement est le contexte : les paires d'entités qui ont le même contexte partagent la même relation. TextRunner proposé par (Banko et al. 2007) permet d'extraire des relations à partir du Web (117 millions de pages). Ils entraînent un classifieur bayésien naif sur un petit corpus analysé syntaxiquement, ce classifieur permet alors de classer toute paire candidates extraite du Web.

Méthodes distributionnelles

Inspirés des travaux de Harris [Harris, 1991], les approches fondées sur la construction de modèles distributionnels consistent à construire une représentation vectorielle des mots à partir de l'analyse de leurs contextes dans de larges corpus de textes [Turney et al., 2010, Lenci, 2008]. Ces méthodes sont basées sur un principe élémentaire selon lequel la proximité sémantique entre deux mots peut être déduite de la proximité de leur distribution. Elles ont bénéficié de la disponibilité de quantités massives de données textuelles et d'une puissance de calcul accrue, permettant l'application de ces méthodes à grande échelle. L'inconvénient de ces méthodes vis-à-vis de notre objectif est que cette mesure de proximité sémantique, très robuste, est aussi très frustre et ne permet pas de distinguer les différentes relations qui peuvent associer deux mots (synonymie, hyperonymie, antonymie) [Fabre and Lenci, 2015]. Certains travaux visent à développer des méthodes pour discriminer les différentes relations lexicales disponibles dans les clusters mis au jour [Santus et al., 2014]. Parmi eux, certains se sont intéressés spécifiquement à la relation d'hyperonymie. Plusieurs pistes ont été définies, consistant à filtrer les résultats d'une approche par patrons par du clustering distributionnel [Schropp et al., 2013], ou à développer des mesures distributionnelles asymétriques pour implémenter le principe d'un recouvrement des contextes distributionnels du terme générique par le terme spécifique [Lenci and Benotto, 2012a]. Bien que ces méthodes obtiennent une couverture élevée si on les comparent aux approches par patrons, leurs scores de précision sont inférieurs car il reste très difficile de déterminer la nature exacte de la relation sémantique entre les termes.

2.3 Discussion

La présentation des travaux visant l'extraction de la relation d'hyperonymie à partir de textes montre la diversité des techniques à notre disposition. Les méthodes à base de règles ont montré leur efficacité pour l'extraction d'informations dans des domaines techniques, bien circonscrits, dans lesquels il est possible d'énumérer de façon efficace les patrons les plus fréquemment utilisés pour exprimer la relation visée [Lefever et al., 2014]. A l'inverse, les approches par apprentissage fournissent des solutions pour faire émerger l'information d'une masse de textes hétérogènes, sur lesquels on ne dispose pas de connaissance experte. Elles offrent une alternative aux approches par patrons dont les performances sont généralement limitées en termes de rappel ; mais leurs résultats sont généralement bruités et nécessitent de mettre en place des procédures de filtrage. Notre contribution consiste à la fois à clarifier les conditions optimales d'utilisation de ces méthodes, et à concevoir des méthodes pour les associer de façon efficace.

3 Une approche par patrons pour Wikipedia

3.1 Introduction

La revue des travaux que nous avons proposée dans le chapitre précédent montre clairement les apports et les limites de l'approche par patrons. Néanmoins, ses performances sont largement dépendantes des conditions d'acquisition des relations sémantiques, et en particulier de la nature du matériau textuel utilisé. Dans cette première série d'expérimentations, nous avons mis en œuvre cette première méthode avec deux objectifs :

— l'appliquer à un corpus, Wikipedia, qui diffère des corpus spécialisés habituellement utilisés dans les études antérieures, particulièrement sur le français,
— comparer ses performances sur plusieurs types de textes présentant des niveaux de structuration distincts,
— mettre au point une méthode qui servira à la fois de baseline, et de méthode complémentaire qui sera combinée à des méthodes par apprentissage dans les étapes suivantes.

3.2 Problématique

Wikipedia est une encyclopédie en ligne riche en contenu textuel. Une des particularités de cette ressource réside dans la diversité des langues de publication. Pour l'anglais, plusieurs travaux ont été réalisés pour extraire des relations d'hyperonymie entre concepts ou classes, ciblant particulièrement la composante *Définition* des pages Wikipedia. De fait, cette composante textuelle est particulièrement adaptée à l'application de méthodes d'extraction basées sur les patrons lexico-syntaxiques, car sa structure est très régulière, ce qui facilite l'application des patrons du type de ceux définis par Hearst. Ce principe pourrait être étendu à d'autres parties des pages Wikipedia, qui présentent des caractéristiques semblables (comme les parties comportant des énumérations). D'autre part, Wikipédia contient des types de pages spécifiques, très structurées, du fait de l'utilisation d'un template de rédaction exigé par la communauté. Parmi ces pages spécifiques figurent les pages appelées *pages de désambigüisation*, sur lesquelles nous avons choisi de focaliser notre première expérimentation.

Dans ce chapitre, nous décrivons une démarche qui vise à évaluer la performance des patrons lexico-syntaxiques sur (i) les paragraphes non structurés des pages Wikipedia et (ii) sur les pages de désambiguïsation, en partant de deux bases de patrons existantes que nous décrivons dans la section suivante.

3.3 Les ressources de patrons CAMELEON et MAR-REL

La création de patrons est une étape importante dans la démarche d'extraction de relations par patrons. C'est une tâche considérée comme coûteuse et fastidieuse à cause de sa complexité. Elle suppose de mettre en œuvre une analyse approfondie du corpus afin de caractériser toutes les réalisations langagières associées à la relation cible. D'autre part, un patron est un objet complexe car il peut être composé d'éléments lexicaux, grammaticaux ou sémantiques [Auger and Barrière, 2008]. Concernant la langue française, deux ressources de patrons ont été proposées dans les communautés du TAL, de la linguistique et de l'ingénierie des connaissances, à savoir : Caméléon et MAR-REL. Dans cette section, nous décrivons ces deux ressources que nous avons utilisées dans nos expérimentations sur le corpus Wikipédia.

3.3.1 La base CAMELEON

L'outil CAMELEON est un logiciel de définition et de projection de patrons réalisé au sein de l'équipe MELODI [Séguéla, 2001]. Cet outil comporte une base de patrons lexico-syntaxiques relatifs à différentes relations sémantiques comme l'hyperonymie, la méronymie, la causalité. Ces patrons sont définis de manière générique, et sont supposés être adaptés aux particularités du domaine ou du corpus étudié. La première expérimentation que nous avons réalisée, et qui est décrite dans l'article qui clôt ce chapitre, a utilisé les patrons d'hyperonymie génériques issus de cette base.

On distingue en effet les patrons génériques, "prêts à l'emploi", qui sont fournis à l'utilisateur de CAMELEON par défaut, et les patrons spécifiques, que l'utilisateur peut définir en adéquation avec les données de son étude. Ces patrons génériques sont au nombre de 70, parmi lesquels figurent 35 patrons marqueurs de l'hyperonymie. L'ensemble des patrons sont disponibles en annexe de l'article [Jacques and Aussenac-Gilles, 2006]. En voici deux exemples qui permettent de repérer une relation CY EST-UN CX entre deux concepts CY et CX respectivements dénnotés par les termes Y et X. Le premier patron permet d'extraire les occurrences du type "Les Occidentaux, et notamment les Américains".

```
X_,_Adv_Y 0-100
NOM|ADJ PUN.*">^,$" ADV">^(notamment|notablement|spécialement
|particulièrement|surtout)$" PRP:det|DET.*
```

Le second permet d'extraire les occurrences du type "juillet, le mois le plus chaud".

```
Y_virgule_le_X_le_plus 0-100
NOM|ADJ PUN.*">^,$" DET:ART">^le$" {NOM|ADJ|ADV|DET.*|PRP.*|KON}*
 DET:ART">^le$" ">^(plus|moins)$" {KON|PRP.*|NOM|DET.*|ADV|ADJ}*
 PUN. *">^,$"
```

Afin de réutiliser ces patrons, les traitements suivants ont été réalisés :

1. Adaptation des étiquettes aux catégories grammaticales de l'analyseur lexico-syntaxiques que nus utilison ;
2. Caractérisation des arguments en relation ;
3. Transcription de ces patrons dans le langage JAPE, de manière à qu'ils soient modulaires et utilisables dans la plate-forme Gate.

Ces patrons CAMELEON transcrits en JAPE ont constitué la base de la première expérimentation que nous avons menée sur les pages Wikipedia.

3.3.2 la base MAR-REL

La base MAR-REL est une liste de candidats-marqueurs conçus pour l'anglais et le français, pour trois relations : l'hyperonymie, la méronymie et la cause. La précision de ces patrons a été évaluée sur des corpus variés du point de vue du domaine et du genre textuel.

Cette ressource [1] a été réalisée dans le cadre du projet CRISTAL (Contextes Riches en connaissances pour la TrAduction terminoLogique), dont les partenaires étaient le LI2N (Laboratoire d'Informatique de Nantes), le laboratoire CLLE, la société Lingua et Machina et la Faculté de Traduction et d'Interprétation de l'Université de Genève. La constitution des candidats-marqueurs et l'analyse de leur fonctionnement en corpus a fait l'objet de la thèse de Luce Lefeuvre [Lefeuvre, 2017]. Cette ressource a été constituée par compilation de patrons antérieurement définis, augmentés de nouveaux patrons. Les marqueurs de cause forment l'essentiel de la ressource (316), complétés par 99 marqueurs de méronymie et 37 marqueurs d'hyperonymie. Les patrons sont regroupés par type. Concernant l'hyperonymie, 6 types sont identifiés : les marqueurs attributifs, appositifs, coordonnés, d'inclusion, d'exemplification, et les marqueurs avec Utiliser.

Voici un exemple de patron attributif, codé au format UIMA.

```
(NominalPhrase{-> MARK(Hyponym)}
Token.lemma == "être"
Token[0, 3]?
Token.partOfSpeech == "Det"
NominalPhrase{-> MARK(Hyperonym)}
Token.lemma == "très"
Token.partOfSpeech == "Adj")
{-> CREATE(Hyperonymy, "category" = "Structures attributives")};
```

Ces deux ressources, CAMELEON et MAR-REL, présentent une intersection importante. Dans les expérimentations ultérieures, nous avons dans un premier temps comparé leurs apports respectifs, puis opté pour une fusion des deux ressources.

1. La ressource est diffusée sous la licence Creative Commons BY-SA 3.0. Elle est disponible sur le site REDAC du laboratoire CLLE : http ://redac.univ-tlse2.fr/misc/mar-rel_fr.html

3.3.3 Comparatif de différentes ressources de patrons

Les taux de rappel et de précision obtenus dans un premier temps à partir des patrons CAMELEON sur une sélection de pages spécifiques, les pages d'homonymie, étaient respectivement de 0.3 et 0.58. Ces faibles taux s'expliquent par le fait que les pages d'homonymie bénéficient d'une mise en forme spécifique (présence de marqueurs typographiques qui suppléent certaines marques lexicales). Or, ces particularités typographiques ne sont pas prise en compte dans les patrons de la base générique de CAMELEON. Nous avons alors défini de nouveaux patrons adaptés à notre corpus, comme en rend compte l'article présenté dans ce chapitre. Il s'agit de 3 patrons permettant de détecter plusieurs cas de structures appositives telles que celles décrites par [Borillo, 1996], comme dans l'exemple *Isaac Babel, écrivain et dramaturge russe (1894-1940).*

Parallèlement, nous avons réalisé d'autres expérimentations en utilisant la ressource MAR-REL. Le tableau ci-dessous résume les résultats obtenus en terme de rappel et précision sur le corpus homonymie de Wikipedia, dans deux cas de figure : utilisation seule de la ressource, utilisation combinée avec les patrons spécifiquement forgés pour ajuster le traitement aux spécificités du corpus.

	P	R
CAMELEON	0 .58	0.30
MAR-REL	0.61	0.47
CAMELEON& Homonymie	0.75	0.68
MAR-REL & Homonymie	**0.77**	**0.71**

Table 3.1 – Comparatif de différents patrons

On constate que les performances obtenues avec les bases MAR-REL et CAMELEON sont similaires, avec un très léger avantage à la base MAR-REL. Ce sont les patrons spécifiques qui améliorent nettement les performances, ce qui confirme les choix méthodologiques des deux équipes ayant mis au point ces bases depatrons, qui suggèrent de définir un jeu de patrons spécifiques, plus ou moins adaptés des patrons génériques, pour chaque nouveau corpus.

L'évaluation de l'expérimentation a été réalisée sur la base de l'annotation de 30 pages Wikipedia sélectionnées aléatoirement à partir du corpus de pages d'homonymie.

3.4 Bilan de la première expérimentation

La première expérimentation qui a donné lieu à publication a consisté à tester la méthode par patrons en appliquant les patrons CAMELEON, augmentés de patrons spécifiques, au traitement d'un sous-corpus particulier, issu de la collection des pages d'homonymie de Wikipedia. Les apports et les limites de cette première approche sont décrits dans l'article reproduit à la fin de ce chapitre.

Ghamnia, A. (2016). Extraction de relations d'hyperonymie à partir de Wikipédia. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2016, volume 3 : RECITAL*, Paris, France, 40-51.

Cette première étape a permis de mettre au point un ensemble de briques de traitement :

— la procédure d'annotation manuelle, qui sera reproduite aux étapes suivantes pour permettre d'évaluer les performances des extracteurs ; de plus les pages d'homonymie annotées pour cette étude seront réutilisées opour les évéluations futures ;
— l'extraction des termes en corpus, qui combine plusieurs outils (taggeur, extracteur) en une chaîne de traitement ; ici encore, ces processus serint repris pour tester d'autres types d'extracteurs ;
— la définition des marqueurs, et leur opérationnalisation via la plate-forme Gate.

Notre expérimenation a par ailleurs permis de montrer que l'extraction de relations à partir du contenu textuel constitue un apport réel pour les deux bases de connaissances visées (DBPedia et BabelNet) : un tiers environ des relations extraites ne sont pas déjà répertoriées dans ces bases de connaissances. L'objectif applicatif de notre démarche se trouve donc conforté.

En retour, cette expérimentation a révélé les limites d'une approche exclusivement fondée sur les patrons, malgré la restriction du traitement à un sous-ensemble de textes à structure régulière. On remarque principalement que les performances de l'extracteur sont affectées par des patrons bruités. Elles pourraient être améliorées en affinant la base des patrons pour la rentre plus adaptée aux textes à traiter, mais cette solution n'est pas envisageable pour traiter l'ensemble du corpus

Wikipedia, qui constitue l'étape ultérieure. Cette première étape a donc mis en évidence l'intérêt de tester une approche par apprentissage pour apprendre des extracteurs spécifiques au corpus.

Extraction de relations d'hyperonymie à partir de Wikipédia

Adel Ghamnia[1]
(1) IRIT, Avenue de l'étudiant, 31400 Toulouse, France
adel.ghamnia@irit.fr

RÉSUMÉ
Ce travail contribue à montrer l'intérêt d'exploiter la structure des documents accessibles sur le Web pour enrichir des bases de connaissances sémantiques. En effet, ces bases de connaissances jouent un rôle clé dans de nombreuses applications du TAL, Web sémantique, recherche d'information, aide au diagnostic, etc. Dans ce contexte, nous nous sommes intéressés ici à l'identification des relations d'hyperonymie présentes dans les pages de désambiguïsation de Wikipédia. Un extracteur de relations d'hyperonymie dédié à ce type de page et basé sur des patrons lexico-syntaxiques a été conçu, développé et évalué. Les résultats obtenus indiquent une précision de 0.68 et un rappel de 0.75 pour les patrons que nous avons définis, et un taux d'enrichissement de 33% pour les deux ressources sémantiques BabelNet et DBPédia.

ABSTRACT
Hypernym extraction from Wikipédia

The volume of available documents on the Web continues to increase, the texts contained in these documents are rich information describing concepts and relationships between concepts specific to a particular field. In this paper, we propose and exploit an hypernymy extractor based on lexico-syntactic patterns designed for Wikipedia semi-structured pages, especially the *disambiguation pages*, to enrich a knowledge base as BabelNet and DBPedia. The results show a precision of 0.68 and a recall of 0.75 for the patterns that we have defined, and an enrichment rate up to 33% for both BabelNet and DBPédia semantic resources.

MOTS-CLÉS : Extraction de relations d'hyperonymie, Base de connaissances, Patrons morpho-syntaxiques.

KEYWORDS: Hypernym extraction, Knowledge Base, morpho-syntactic patterns.

1 Introduction

L'objectif de notre travail est l'enrichissement de bases de connaissances sémantiques de type BabelNet (Navigli & Ponzetto, 2012) ou DBPédia (Lehmann *et al.*, 2014) à partir des informations contenues dans des documents textuels semi-structurés. Ces bases de connaissances jouent aujourd'hui un rôle clé dans de nombreuses applications du TAL, et leur alimentation constitue donc un enjeu important afin de rendre disponibles des informations lexico-sémantiques multilingues à large échelle. A l'heure actuelle, la construction de ces réseaux se fonde principalement sur des ressources existantes telles WordNet ou sur l'exploitation de la partie structurée des documents encyclopédiques de

Wikipédia [1]. Ainsi, des extracteurs dédiés se focalisent sur les infobox, les catégories, ou les liens définis dans les pages Wikipédia (Morsey *et al.*, 2012; Lehmann *et al.*, 2014). Les contenus textuels des documents, riches en information décrivant des concepts et des relations entre ces concepts, mais plus difficilement accessibles, sont généralement sous-exploités.

Différentes méthodes ont cependant été définies pour extraire à partir des textes des informations (termes et relations sémantiques entre termes) susceptibles d'alimenter ces bases de connaissances. Ces travaux utilisent généralement des extracteurs de termes et font appel à des techniques fondées sur l'application de patrons morpho-syntaxiques dans la lignée de (Hearst, 1992), sur le principe de proximité distributionnelle (Lenci & Benotto, 2012), ou sur l'exploitation de structures textuelles spécifiques, par exemple les définitions (Malaisé *et al.*, 2004) ou les structures énumératives (Fauconnier & Kamel, 2015).

Notre travail de recherche vise à enrichir le Web des données pour le français en mettant en oeuvre de façon combinée plusieurs méthodes d'extraction de termes et de relations entre termes à partir des textes, pour l'acquisition de différents types de relations sémantiques, en premier lieu l'hyperonymie et la méronymie. Comme il a été montré par exemple par (Schropp *et al.*, 2013), la combinaison de plusieurs approches est une piste intéressante pour tirer parti de la multiplicité des indices textuels signalant une relation sémantique, et dépasser les limites identifiées pour chaque méthode. Notre approche s'appuiera sur un corpus issu de Wikipédia en français, dont les articles ont la particularité de combiner différents niveaux de structuration de l'information textuelle.

Nous présentons dans cet article une étape préliminaire de ce travail, qui vise à tester la démarche à partir d'un premier cas d'étude. Notre premier objectif est de déterminer la plus-value potentielle de l'extraction de relations à partir de textes, en évaluant l'apport d'une première expérience d'extraction pour l'alimentation des bases DBPédia et BabelNet. Dans le cadre de cet article, nous nous focalisons sur l'extraction de la relation d'hyperonymie à partir de certains articles de Wikipédia appelés pages de désambiguïsation, et nous nous limitons à la démarche classique d'extraction par patrons lexico-syntaxiques. Nous amorçons notre approche en choisissant un cas de figure favorable, puisque les pages de désambiguïsation de Wikipédia sont riches en entités nommées et en relations d'hyperonymie exprimées dans des structures textuelles contraintes, généralement des énumérations, et normées par les consignes de la charte de rédaction. Cela nous permet de concevoir une liste de patrons lexico-syntaxiques adaptés à ce type de pages, qui couvrent à la fois la relation sémantique et les arguments de cette relation, en tirant parti de la structure particulière de ces pages.

Dans ce qui suit, nous proposons tout d'abord un état de l'art des méthodes d'extraction des relations d'hyperonymie à partir de textes. Nous présentons ensuite notre corpus et les annotations réalisées. La troisième partie décrit la méthode d'extraction de relations, avant une section consacrée à l'évaluation de la méthode, à la fois intrinsèque (performance des patrons) et extrinsèque (apport pour l'alimentation des bases sémantiques).

2 Travaux précédents

Dans cette partie, nous faisons une courte synthèse des travaux réalisés sur l'extraction de la relation d'hyperonymie, et plus particulièrement sur l'extraction de relations à partir de Wikipédia.

1. fr.wikipedia.org

2.1 Extraction de relations d'hyperonymie

De nombreux travaux ont été consacrés à l'extraction automatique de la relation d'hyperonymie. On peut les organiser en deux grands types de démarches. Une première série de travaux est inspirée du travail pionnier de Hearst (1992), qui a montré que la relation d'hyperonymie est directement signalée dans les textes dans des constructions régulières, et peut être extraite par la projection de patrons lexico-syntaxiques reliant deux termes. Ce travail a inspiré de nombreux travaux qui ont progressivement intégré des méthodes d'apprentissage pour limiter le coût de construction des patrons (Morin & Jacquemin, 2004; Snow *et al.*, 2004; Pantel & Pennacchiotti, 2006). Le travail de (Panchenko *et al.*, 2013) a réadapté et évalué les patrons de Hearst sur des gros corpus en français. (Panchenko *et al.*, 2016; Bordea *et al.*, 2015) ont utilisé une méthode d'extraction de relations basée sur des patrons pour reconstruire des taxonomies existantes à partir d'une liste finie de termes. Ces travaux ont montré que l'approche par patrons produit des résultats généralement satisfaisants du point de vue de la précision. Néanmoins, comme ils n'exploitent qu'une partie infime de l'information textuelle, leur couverture est extrêmement réduite.

Un deuxième type d'approche est fondé sur l'hypothèse distributionnelle, qui consiste à établir une relation de proximité entre deux unités lexicales qui présentent des propriétés distributionnelles semblables. On retrouve dans cette catégorie les travaux de (Caraballo, 2001) ou (Van Der Plas *et al.*, 2005). On sait que l'approche distributionnelle a pour effet de mettre au jour une relation de proximité sémantique au sens large, mêlant indifféremment des relations d'hyperonymie, synonymie, co-hyponymie, ou des relations de proximité plus lâche. Il s'agit donc d'identifier au sein des espaces distributionnels des relations de voisinage qui présentent des propriétés spécifiques, par exemple en se fondant, comme le font Lenci & Benotto (2012), sur une hypothèse d'inclusion distributionnelle, les traits distributionnels de l'hyponyme étant inclus dans ceux de l'hyperonyme.

D'autres travaux explorent encore des voies complémentaires, en tirant parti de caractéristiques structurelles ou dispositionnelles des textes, afin de repérer des zones denses en relations sémantiques, en particulier les structures énumératives (Fauconnier & Kamel, 2015). Une technique plus simple, dite d'inclusion lexicale, est également utilisée pour tirer parti de la structuration interne des textes (Lefever *et al.*, 2014), la tête lexicale d'un terme complexe pouvant être considérée comme l'hyperonyme du terme complet.

Ces approches sont complémentaires : les unes s'intéressent aux segments dans lesquels les deux termes cooccurrent dans des contextes spécifiques, d'autres au contraire tirent parti de l'ensemble des occurrences des termes dans le corpus, ou de leurs propriétés formelles. La combinaison des différents types d'approches s'avère de fait prometteuse, comme démontré par (Schropp *et al.*, 2013).

2.2 Extraction de relations à partir de Wikipédia

DBPédia est une base de connaissances qui contient des concepts et relations extraits de Wikipédia. Morsey *et al.* (2012) ont développé 19 extracteurs, chacun étant dédié au traitement d'un type particulier d'information structurée au sein des pages Wikipédia : infobox, catégorie, lien, image, etc. La figure 1 montre la façon dont les pages Wikipédia font coexister des structures directement accessibles par ce type d'extracteurs (par exemple, les liens entre entités), et du contenu textuel, dont l'essentiel n'est pas exploité.

Ainsi, les travaux d'extraction de relations d'hyperonymie à partir de Wikipédia ciblent particuliè-

FIGURE 1 – Une page Wikipédia

rement les parties structurées. Par exemple, Suchanek *et al.* (2007) ont utilisé les catégories, qui constituent le système de classement thématique de Wikipédia ; Kazama & Torisawa (2007) ont exploité la partie définition ; enfin Sumida & Torisawa (2008) se sont intéressés aux menus, qui offrent un moyen d'accéder à la hiérarchie des concepts.

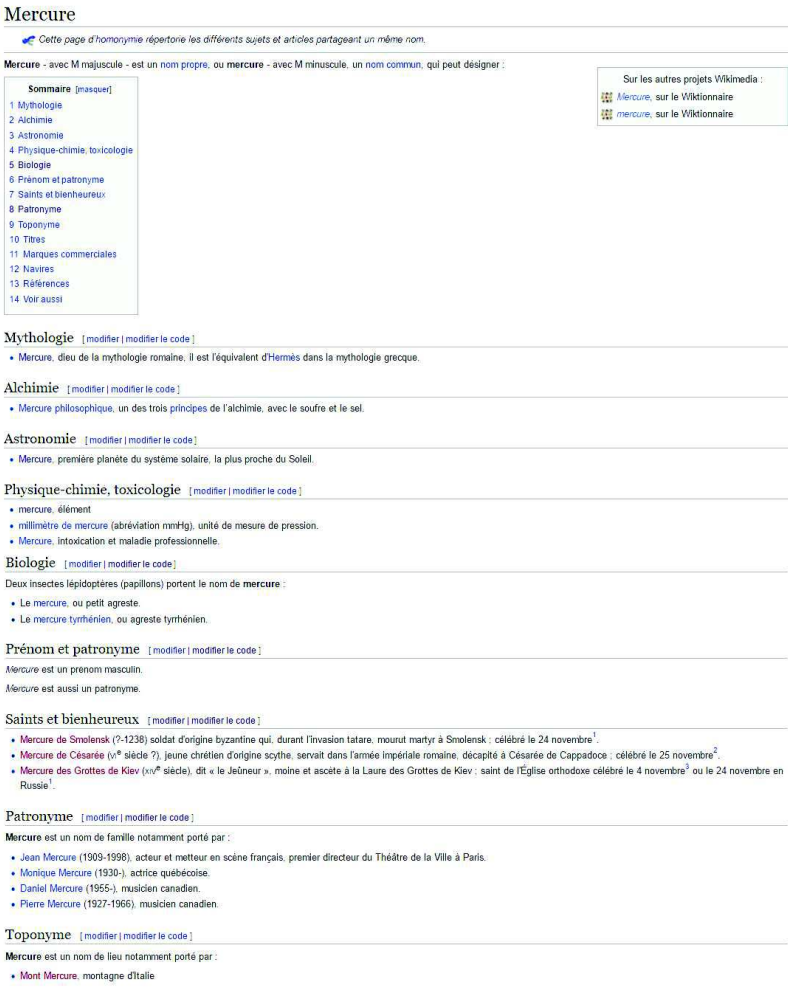
3 Corpus et données d'annotation

Nous avons constitué un corpus à partir du *dump* de la version française de l'encyclopédie Wikipédia. Nous avons choisi de nous focaliser dans ce travail sur un type particulier de pages, appelées pages de désambiguïsation (appelée aussi page d'homonymie). Ces pages listent les articles dont le titre est ambigu, et donnent une définition de toutes les acceptions recensées. Ces pages sont riches en relations d'hyperonymie et ont une structure régulière. Par exemple, la page d'homonymie *Mercure* cite plusieurs articles, parmi lesquels :

— le dieu romain Mercure ;
— la planète Mercure ;
— l'élément chimique mercure ;

Une page de désambiguïsation est structurée en plusieurs rubriques correspondant à différents types d'homonymie. Deux d'entre elles, consacrées au recensement des patronymes et des toponymes, ont une structure régulière, comme illustré dans la figure 2. Ces deux rubriques doivent être rédigées selon un *template* proposé par Wikipédia. Cette normalisation est généralement adoptée par le rédacteur, ce qui nous a permis de définir des patrons spécifiques pour le repérage des relations d'hyperonymie.

Afin d'évaluer notre approche, nous avons constitué un sous-corpus composé de 30 pages de désambiguïsation, tirées aléatoirement du corpus. Le tableau 1 détaille les 30 pages utilisées et le nombre



Mercure

Cette page d'homonymie répertorie les différents sujets et articles partageant un même nom.

Mercure - avec M majuscule - est un nom propre, ou **mercure** - avec M minuscule, un nom commun, qui peut désigner :

Sommaire [masquer]
1 Mythologie
2 Alchimie
3 Astronomie
4 Physique-chimie, toxicologie
5 Biologie
6 Prénom et patronyme
7 Saints et bienheureux
8 Patronyme
9 Toponyme
10 Titres
11 Marques commerciales
12 Navires
13 Références
14 Voir aussi

Sur les autres projets Wikimedia :
Mercure, sur le Wiktionnaire
mercure, sur le Wiktionnaire

Mythologie [modifier | modifier le code]
- Mercure, dieu de la mythologie romaine, il est l'équivalent d'Hermès dans la mythologie grecque.

Alchimie [modifier | modifier le code]
- Mercure philosophique, un des trois principes de l'alchimie, avec le soufre et le sel.

Astronomie [modifier | modifier le code]
- Mercure, première planète du système solaire, la plus proche du Soleil.

Physique-chimie, toxicologie [modifier | modifier le code]
- mercure, élément
- millimètre de mercure (abréviation mmHg), unité de mesure de pression.
- Mercure, intoxication et maladie professionnelle.

Biologie [modifier | modifier le code]

Deux insectes lépidoptères (papillons) portent le nom de **mercure** :
- Le mercure, ou petit agreste.
- Le mercure tyrrhénien, ou agreste tyrrhénien.

Prénom et patronyme [modifier | modifier le code]

Mercure est un prenom masculin.

Mercure est aussi un patronyme.

Saints et bienheureux [modifier | modifier le code]
- Mercure de Smolensk (?-1238) soldat d'origine byzantine qui, durant l'invasion tatare, mourut martyr à Smolensk ; célébré le 24 novembre[1].
- Mercure de Césarée (VI^e siècle ?), jeune chrétien d'origine scythe, servait dans l'armée impérialale romaine, décapité à Césarée de Cappadoce ; célébré le 25 novembre[2].
- Mercure des Grottes de Kiev (XIV^e siècle), dit « le Jeûneur », moine et ascète à la Laure des Grottes de Kiev ; saint de l'Église orthodoxe célébré le 4 novembre[3] ou le 24 novembre en Russie[1].

Patronyme [modifier | modifier le code]

Mercure est un nom de famille notamment porté par :
- Jean Mercure (1909-1998), acteur et metteur en scène français, premier directeur du Théâtre de la Ville à Paris.
- Monique Mercure (1930-), actrice québécoise.
- Daniel Mercure (1955-), musicien canadien.
- Pierre Mercure (1927-1966), musicien canadien.

Toponyme [modifier | modifier le code]

Mercure est un nom de lieu notamment porté par :
- Mont Mercure, montagne d'Italie

FIGURE 2 – Une page de désambiguïsation

de relations d'hyperonymie qu'elles contiennent. Le corpus a subi une phase de nettoyage pour extraire le texte à partir de la version XML. Nous avons ensuite utilisé TreeTagger [2] pour l'étiquetage morpho-syntaxique. Enfin, nous avons utilisé l'extracteur de termes YaTeA (Aubin & Hamon, 2006) pour identifier les syntagmes nominaux qui occupent les positions d'arguments dans la relation d'hyperonymie. La dernière étape du processus consiste à transformer le format de sortie de TreeTagger au format requis par l'environnement d'ingéniere linguistique Gate, afin d'exécuter les patrons lexico-syntaxiques.

L'évaluation des systèmes d'extraction de la relation d'hyperonymie se fonde généralement sur des ressources lexicales telles que WordNet. En l'absence d'une ressource aussi complète pour le français, nous avons opté pour la création d'une annotation de référence, en marquant toutes les occurrences de la relation d'hyperonymie dans notre corpus. Ce sous-corpus utilisé contient 30 pages de désambiguïsation, 9718 tokens et 553 relations d'hyperonymie annotées manuellement.

2. http ://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger

Nom de la page	Nbre de relations d'hyperonymie annotées
Timbre	20
Souris	26
Samurai	4
Renaissance	26
Produit	37
Prairial	6
Pluton	9
Pentagone	8
Opera	19
Morville	12
Montreuil	41
Matrice	33
Magma	9
Lumen	10
Louis Philippe	13
Lincoln	61
Kikai	5
Henri Giraud	8
Gaulois	8
Gandhi	25
Divergence	10
Coulombs	3
Cornet	30
Colombier	34
Calypso	23
Champollion	3
Apache	14
Analyse	8
Ampoule	10
Columbia	38

TABLE 1 – Le sous-corpus utilisé

4 L'extracteur d'hyperonymie

L'extracteur d'hyperonymie que nous proposons pour les pages d'homonymie est basé sur la définition de patrons lexico-syntaxiques, constitués d'un ensemble d'une dizaine de patrons proposés par (Jacques & Aussenac-Gilles, 2006) et augmentés de patrons spécifiques visant à capter les spécificités des rubriques patronymes et toponymes des pages de désambiguïsation. Voici à titre d'exemple un extrait de la rubrique *Patronymes* de la page de désambiguïsation *Babel* :

1. Louis Babel, prêtre-missionnaire oblat et explorateur du Nouveau-Québec (1826-1912).
2. Isaac Babel, écrivain et dramaturge russe (1894-1940).
3. Ryan Babel, joueur de football batave (1986-).
4. Roger Viry-Babel (1945-2006), universitaire et cinéaste français.

Le tableau 2 regroupe les relations qui peuvent être extraites de cette rubrique et les place en regard de celles qui existent actuellement dans DBPédia pour les entités correspondantes, à savoir *Louis Babel*, *Isaac Babel*, *Ryan Babel*, et *Roger Viry-Babel*. L'ontologie de DBPédia modélise les relations d'hyperonymie par la propriété RDF *rdf :type* pour les relations entre une instance et sa classe, et la propriété RDF *rdf :subClassOf* pour les relations entre classes. Cette comparaison permet de constater que 5 des 12 relations décrites dans le texte ne sont actuellement pas recensées dans DBPédia.

Phrase	Relations	Relations dans DBPédia
(1)	*Hyp*(Louis Babel, Prêtre-missionnaire) *Hyp*(Louis Babel, Explorateur)	*rdf :type*(Louis Babel, Agent) *rdf :type*(Louis Babel, Personne)
(2)	*Hyp*(Issac Babel, Ecrivain) *Hyp*(Issac Babel, Dramaturge)	*rdf :type*(Issac Babel, Agent) *rdf :type*(Issac Babel, Artiste) *rdf :type*(Issac Babel, Personne) *rdf :type*(Issac Babel, Ecrivain)
(3)	*Hyp*(Ryan Babel, Joueur de football)	*rdf :type*(Ryan Babel, Agent) *rdf :type*(Ryan Babel, Athlete) *rdf :type*(Ryan Babel, Personne) *rdf :type*(Ryan Babel, Joueur de football)
(4)	*Hyp*(Roger Viry-Babel, Universitaire) *Hyp*(Roger Viry-Babel,Cinéaste)	*rdf :type*(Roger Viry-Babel, Agent) *rdf :type*(Roger Viry-Babel, Personne)

TABLE 2 – Exemples de relations et comparaison avec DBPédia

Des caractéristiques semblables concernent également la rubrique des toponymes, ce qui nous a amené à proposer des patrons pour extraire plus spécifiquement les relations d'hyperonymie présentes dans ces deux rubriques. Ces patrons sont décrits à l'aide des Expressions Régulières (ER) suivantes :

1. NP '(' { ({NUM - (NUM | ' ?')} | NUM) } ')' , NP {, NP}* {(et|ou) NP} ?
2. NP (,| :) NP {, NP}* {(et|ou) NP} ?
3. NP '(' NP {, NP}* {(et|ou) NP} ? ')'

NP (pour *Noun Phrase*) correspond aux traces linguistiques des arguments de la relation d'hyperonymie (l'hyperonyme et l'hyponyme), Nous avons utilisé l'extracteur de terme YaTeA pour annoter les

syntagmes nominaux en prenant en compte la proposition la plus longue que propose cet extracteur. Par exemple pour la phrase *Une vache ayant quatre sabots est un animal*, YaTeA propose 2 termes : *une vache* et *une vache ayant quatre sabots*. On considère comme argument la seconde proposition correspondant au terme le plus long.

Le tableau 3 présente des exemples des relations extraites avec ces patrons.

Patron	Texte reconnu	Relation extraite
(1)	Sonia Gandhi (1946), présidente du Parti du Congrès	*Hyp*(Sonia Gandhi, présidente du Parti du Congrès)
(2)	Gopalkrishna Gandhi, homme politique	*Hyp*(Gopalkrishna Gandhi, homme politique)
(3)	atelier (local, espace)	*Hyp*(atelier, local)
Hearst	Gandhi est un film	*Hyp*(Gandhi, film)

TABLE 3 – Exemples de relations extraites

Les relations d'hyperonymie présentes dans les pages d'homonymie relèvent de deux types du point de vue de l'ontologie : relation de typage (*is-a* ou *type-of*) et relation d'instance (*instance-of*). Les rubriques Patronymes et Toponymes contiennent généralement des entités nommées pour les lieux et les personnes, donc des relations d'instance. Par ailleurs, les autres rubriques sont riches en relations de typage ; par exemple *Hyp*(Souris, Dispositif informatique) extraite de la page *Souris* est une relation de typage. Le type des arguments pourrait donc permettre de distinguer ces deux types de relation, ce que nous n'avons pas fait jusqu'à présent. Les patrons proposés ont été définis en analysant 20 pages de désambiguïsation de Wikipédia. Dans la section suivante nous présentons leur évaluation sur un corpus de 30 pages différentes que nous avons présenté dans la section 3.

5 Evaluation et discussion

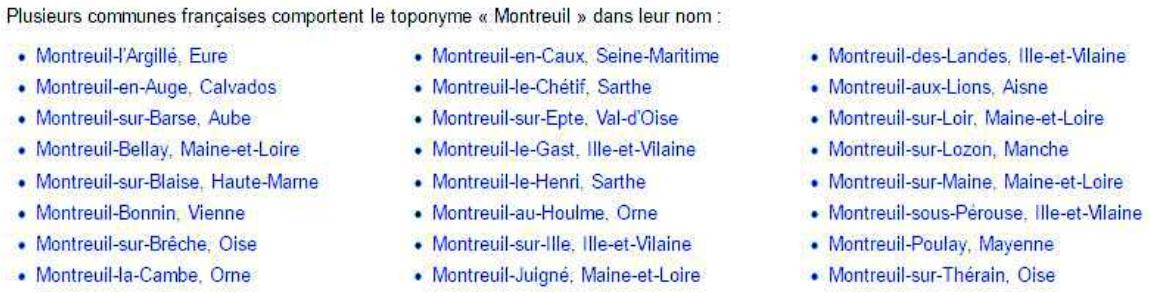
A titre indicatif, nous avons calculé le rappel et la précision par rapport à notre corpus annoté manuellement (tableau 4).

	Rappel	Précision
Corpus	0.75	0.68

TABLE 4 – Evaluation des patrons

Nous remarquons que la précision est inférieure au rappel, ce qui est atypique dans le cas des approches basées sur patrons. Ce résultat confirme que la relation d'hyperonymie s'exprime de façon très régulière dans le corpus particulier que nous avons traité. Néanmoins, certains patrons pourraient être affinés pour éliminer certains faux positifs. C'est en particulier le cas du patron (2), qui génère une quantité importante de bruit malgré la restriction à des termes situés en début de phrase. Par exemple la phrase *Dans la mythologie grecque, Calypso est une nymphe* contient la relation

Hyp(Calypso, nymphe). Cette relation est correctement retrouvée par un patron générique défini par Hearst. Par contre, le patron (2) extrait la relation *Hyp*(mythologie grecque, Calypso), qui est erronée. La présence de listes énumératives verticales dans la rubrique *Patronyme* n'est pas prise en compte par ces patrons. Par exemple, la page de désambiguïsation *Montreuil* contient une grande quantité de relations d'hyperonymie présentes dans une liste énumérative verticale, telle qu'illustrée par la figure 3. Les patrons que nous proposons ne sont pas adaptés au traitement de tels exemples, et devraient donc être complétés. Nous envisageons de définir un modèle d'apprentissage des patrons basé sur la structure de texte en nous appuyant sur l'approche de (Fauconnier & Kamel, 2015) pour le traitement des structures énumératives verticales.


Plusieurs communes françaises comportent le toponyme « Montreuil » dans leur nom :

- Montreuil-l'Argillé, Eure
- Montreuil-en-Auge, Calvados
- Montreuil-sur-Barse, Aube
- Montreuil-Bellay, Maine-et-Loire
- Montreuil-sur-Blaise, Haute-Marne
- Montreuil-Bonnin, Vienne
- Montreuil-sur-Brêche, Oise
- Montreuil-la-Cambe, Orne
- Montreuil-en-Caux, Seine-Maritime
- Montreuil-le-Chétif, Sarthe
- Montreuil-sur-Epte, Val-d'Oise
- Montreuil-le-Gast, Ille-et-Vilaine
- Montreuil-le-Henri, Sarthe
- Montreuil-au-Houlme, Orne
- Montreuil-sur-Ille, Ille-et-Vilaine
- Montreuil-Juigné, Maine-et-Loire
- Montreuil-des-Landes, Ille-et-Vilaine
- Montreuil-aux-Lions, Aisne
- Montreuil-sur-Loir, Maine-et-Loire
- Montreuil-sur-Lozon, Manche
- Montreuil-sur-Maine, Maine-et-Loire
- Montreuil-sous-Pérouse, Ille-et-Vilaine
- Montreuil-Poulay, Mayenne
- Montreuil-sur-Thérain, Oise

FIGURE 3 – Exemple de liste énumérative verticale

Rappelons que l'évalution des patrons eux-mêmes n'était pas l'objectif principal de cette première étape du travail. Notre intention était de commencer à mesurer la capacité de notre démarche à enrichir les ressources sémantiques DBPédia et BabelNet. Pour cela, nous avons comparé le nombre de relations valides repérées par notre approche aux relations présentes dans ces deux ressources et impliquant les mêmes entités. Cette évaluation s'est effectuée en trois étapes :

1. pour chaque relation trouvée, interroger DBPédia et BabelNet pour vérifier l'existence de l'entité hyponyme ;
2. si cet hyponyme existe, récupérer tous les hyperonymes identifiés dans DBPédia et BabelNet ;
3. comparer les hyperonymes trouvés dans DBPédia et BabelNet avec les hyperonymes trouvés par notre approche.

Le tableau 5 montre le pourcentage des relations trouvées avec notre approche et qui ne sont pas présentes dans DBPédia (v19.01.2015) et BabelNet (v3.6). Nous précisons que ces pourcentages concernent les relations valides, donc après élimination des relations fausses trouvées par notre extracteur :

DBPédia	BabelNet
33.49%	33.01%

TABLE 5 – Proportion des relations absentes des ressources

On constate donc que l'application des patrons permettrait de compléter de façon conséquente les ressources disponibles dans DBPédia et BabelNet. Les pourcentages de relations manquantes dans DBPédia et BabelNet sont equivalents. De plus, les relations absentes dans les deux cas sont les

mêmes. On peut faire l'hypothèse que cette proximité tient en partie au fait que les deux ressources sont partiellement construites sur les mêmes principes, à savoir l'exploitation des éléments structurés de Wikipédia. Ce qui expliquerait pourquoi les relations absentes dans les deux ressources sont les mêmes. Ci-dessous quelques exemples de relations d'hyperonymie extraites par notre approche et qui n'existent pas dans DBPédia et BabelNet :

— la souris est le muscle charnu qui tient à l'os du manche d'un gigot : *Hyp*(Souris, Muscle) ;
— Cornet à dés, gobelet servant à mélanger puis jeter les dés : *Hyp*(Cornet à dés, gobelet) ;
— Le Cornet, goguette fondée en 1896 par Georges Courteline : *Hyp*(Le Cornet, goguette) ;
— Charles Joseph Cornet (1879-1914), explorateur et écrivain français : *Hyp*(Charles Joseph Cornet, explorateur), *Hyp*(Charles Joseph Cornet, écrivain).

6 Conclusion

Dans cet article, nous avons présenté notre approche d'extraction de relations d'hyperonymie à partir de Wikipédia pour enrichir DBPédia, en particulier à partir des pages de désambiguïsation. L'évaluation de notre travail a montré que les textes de Wikipédia contiennent aussi des relations d'hyperonymie qui ne sont pas présentes dans les parties structurées. Cette première expérience est une toute première étape par rapport à l'objectif général que nous nous sommes fixé. Rappelons que celui-ci consiste à combiner un ensemble de méthodes d'extraction de relations sémantiques afin d'alimenter des ressources sémantiques du Web des données en français. Nous avons testé notre démarche sur un type de textes spécifique, les pages de désambiguïsation de Wikipédia, en nous limitant à une approche par patrons. Ce premier travail nous a cependant permis de constituer une première chaîne de traitement consistant à prétraiter le *dump* du corpus Wikipédia (étiquetage et extraction de termes), à produire un corpus annoté, à projeter un ensemble de patrons morpho-syntaxiques, et à connecter nos résultats aux ressources BabelNet et DBPédia afin de s'assurer de l'apport de ressources complémentaires issues directement des textes de Wikipédia, et non plus seulement des informations structurées contenues dans ces articles. L'étape suivante va consister à mettre en oeuvre à plus grande échelle l'approche par patrons, en faisant appel à des techniques d'apprentissage, et en considérant cette fois l'intégralité du corpus Wikipédia.

Références

AUBIN S. & HAMON T. (2006). Improving term extraction with terminological resources. In *Advances in Natural Language Processing*, p. 380–387. Springer.

BORDEA G., BUITELAAR P., FARALLI S. & NAVIGLI R. (2015). SemEval-2015 task 17 : Taxonomy Extraction Evaluation (TExEval). *SemEval-2015*, **452**(465), 902.

CARABALLO S. (2001). *Automatic Acquisition of a Hypernym-Labeled Noun Hierarchy from Text. Brown University*. PhD thesis.

FAUCONNIER J.-P. & KAMEL M. (2015). Discovering Hypernymy Relations using Text Layout. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, p. 249–258, Denver, Colorado : Association for Computational Linguistics.

HEARST M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, p. 539–545 : Association for Computational Linguistics.

JACQUES M.-P. & AUSSENAC-GILLES N. (2006). Variabilité des performances des outils de TAL et genre textuel. volume 47, p. 11–32.

KAZAMA J. & TORISAWA K. (2007). Exploiting Wikipedia as external knowledge for named entity recognition. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, p. 698–707.

LEFEVER E., VAN DE KAUTER M. & HOSTE V. (2014). Evaluation of automatic hypernym extraction from technical corpora in English and Dutch. In *9th International Conference on Language Resources and Evaluation (LREC)*, p. 490–497 : European Language Resources Association (ELRA).

LEHMANN J., ISELE R., JAKOB M., JENTZSCH A., KONTOKOSTAS D., MENDES P., HELLMANN S., MORSEY M., VAN KLEEF P., AUER S. & BIZER C. (2014). DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*.

LENCI A. & BENOTTO G. (2012). Identifying hypernyms in distributional semantic spaces. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, p. 75–79 : Association for Computational Linguistics.

MALAISÉ V., ZWEIGENBAUM P. & BACHIMONT B. (2004). Detecting semantic relations between terms in definitions. In *COLING*, p. 55–62.

MORIN E. & JACQUEMIN C. (2004). Automatic acquisition and expansion of hypernym links. *Computers and the Humanities*, **38**(4), 363–396.

MORSEY M., LEHMANN J., AUER S., STADLER C. & HELLMANN S. (2012). DBpedia and the live extraction of structured data from Wikipedia. *Program*, **46**(2), 157–181.

NAVIGLI R. & PONZETTO S. P. (2012). Babelnet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, **193**, 217–250.

PANCHENKO A., FARALLI S., RUPPERT E., REMUS S., NAETS H., FAIRON C., PONZETTO S. P. & BIEMANN C. (2016). Taxi : a Taxonomy Induction Method based on Lexico-Syntactic Patterns, Substrings and Focused Crawling. In *Proceedings of the 10th International Workshop on Semantic Evaluation*.

PANCHENKO A., NAETS H., BROUWERS L., ROMANOV P. & FAIRON C. (2013). Recherche et visualisation de mots sémantiquement liés. *TALN-RÉCITAL 2013*, p. 747–754.

PANTEL P. & PENNACCHIOTTI M. (2006). Espresso : Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, p. 113–120 : Association for Computational Linguistics.

SCHROPP G., LEFEVER E. & HOSTE V. (2013). A Combined Pattern-based and Distributional Approach for Automatic Hypernym Detection in Dutch. In G. ANGELOVA, K. BONTCHEVA & R. MITKOV, Eds., *RANLP*, p. 593–600 : RANLP 2013 Organising Committee / ACL.

SNOW R., JURAFSKY D. & NG A. Y. (2004). Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems 17*.

SUCHANEK F. M., KASNECI G. & WEIKUM G. (2007). Yago : A Core of Semantic Knowledge Unifying WordNet and Wikipedia. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, p. 697–706, New York, NY, USA : ACM.

SUMIDA A. & TORISAWA K. (2008). Hacking wikipedia for Hyponymy Relation Acquisition. In *IJCNLP*, volume 8, p. 883–888.

VAN DER PLAS L., BOUMA G. & MUR J. (2005). Automatic acquisition of lexico-semantic knowledge for QA. In *Proceedings of the IJCNLP workshop on Ontologies and Lexical Resources*, p. 76–84.

4 Approche statistique pour l'extraction de relations

4.1 Problématique

Les méthodes d'extraction de relations par patrons lexico-syntaxiques exploitent les différents indices syntaxiques et linguistiques de la phrase qui contient une relation [Hearst, 1992]. Leur limite majeure réside dans la définition des patrons en fonction de chaque corpus, ce qui est souvent coûteux, chronophage et nécessite une expertise humaine. De plus, il est difficile de caractériser ainsi toutes les occurrences de relation. Les patrons restituent peu d'erreurs mais manquent beaucoup d'expressions de relations lorsqu'elles ne répondent à aucun patron. Comme nous l'avons déjà mentionné dans le chapitre d'état de l'art, les méthodes statistiques sont une solution à ce problème, car il est plus facile de les entraîner sur des phrases rendant compte de la diversité des expressions de relations. Cependant elles nécessitent la définition d'un ensemble d'exemples d'apprentissage qui exigent une phase préalable d'annotation manuelle des corpus à traiter, elle aussi coûteuse en temps pour des experts du corpus. L'apprentissage par supervision distante est une solution qui permet de s'affranchir de cette étape car son principe est d'utiliser une ressource externe pour annoter automatiquement l'ensemble d'exemples d'apprentissage [Min et al., 2013].

Nous décrivons dans ce chapitre ce nouveau scénario d'extraction de relations. Nous étudions la performance de la supervision distante appliquée aux mêmes données que celles traitées dans le précédent chapitre, à savoir les pages de désambiguïsation de Wikipedia ; puis nous étendons l'expérimentation à l'ensemble du corpus Wikipedia.

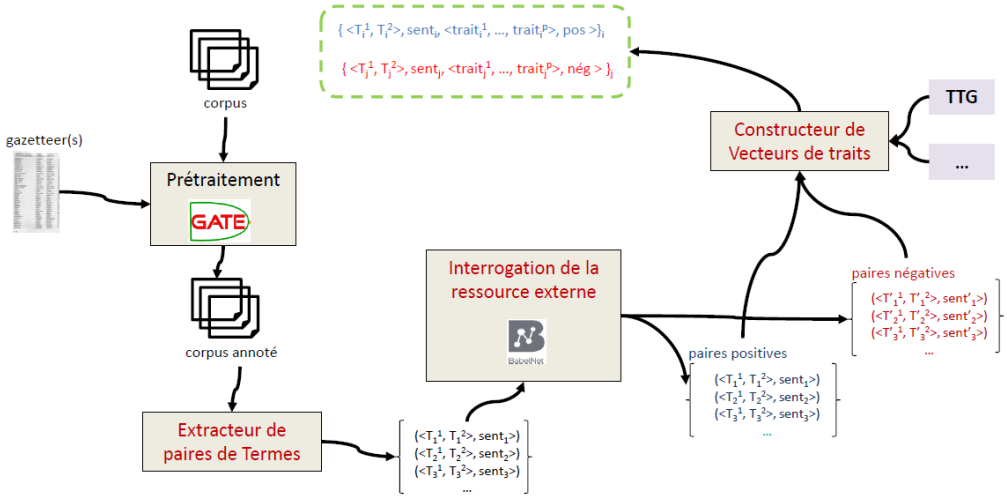
Nous présentons aussi les résultats d'utilisation de différents traits d'apprentissage, et évaluons le taux des relations résultantes qui sont absentes de la base de connaissances DBpedia en français. Concernant la ressource externe que nous utilisons pour la supervision distante, il s'agit de la base de connaissances BabelNet Navigli and Ponzetto [2012], une ressource construite partiellement à partir des pages Wikipedia.

4.2 Extraction par supervision distante

Nous avons choisi de reprendre le principe de la supervision distante proposé par Mintz et al. [2009]. Comme pour tout apprentissage supervisé, il convient de créer un ensemble d'exemples, d'entraîner un modèle statistique sur ces exemples, et d'évaluer le modèle sur un ensemble de test ou par validation croisée. L'originalité de la supervision distante réside dans le fait de construire les exemples automatiquement à partir d'une ressource externe. Dans le cadre de l'extraction de relations, un exemple est une unité textuelle (phrase ou partie de phrase) contenant deux termes ; l'exemple est associé à la relation qui relie ces termes si elle existe, il est négatif si aucune relation ne les relie. La construction d'un exemple consiste à extraire une paire de termes d'une unité textuelle, à la représenter par un ensemble de valeurs de traits préalablement définis, et à lui associer une classe. Dans le cas de la supervision distante, la classe correspond à la relation (si elle existe) qui lie, dans la ressource externe, deux concepts ayant pour labels les deux termes extraits du texte. Une fois entraîné à partir des exemples classés, un algorithme de classification multi-classes permet d'associer une classe à chaque exemple d'un nouveau corpus.

Nous avons adapté cette méthode en nous focalisant sur la relation d'hyperonymie, et en procédant à une classification binaire. La procédure complète est schématisée page suivante en 3 étapes, depuis la construction des exemples jusqu'à l'extraction des relations.

- **Etape 1** : construction des exemples



La construction des exemples enchaîne les traitements suivants :

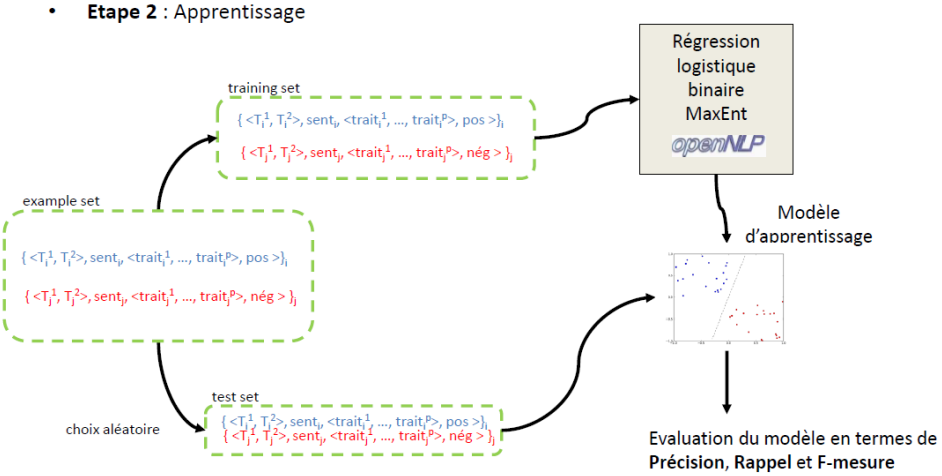
— On effectue une série de pré-traitements sur le corpus : lemmatisation, analyse morpho-syntaxique pour identifier la catégorie grammaticale de chaque mot, élimination de mots vides ; le résultat est un document associé au corpus et contenant les résultats de ces analyses.

— On applique ensuite un extracteur de termes (par exemple l'extracteur Yatea [Aubin and Hamon, 2006]) qui repère les termes présents dans chaque phrase. On dispose alors de toutes les paires possibles par phrase.

— Il s'agit ensuite de retrouver les concepts désignés par ces termes dans BabelNet et de rechercher si ces concepts sont en relation dans cette ressource.

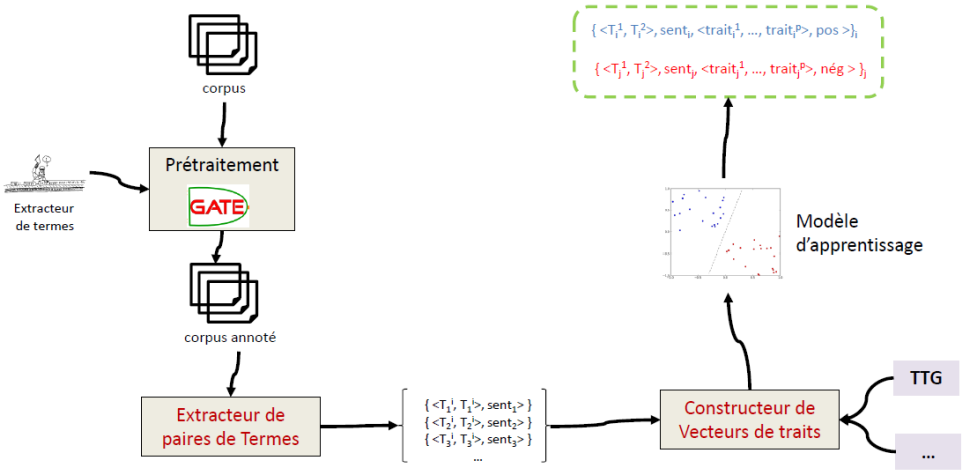
— Pour chaque phrase contenant deux termes associés à des concepts BabelNet, un vecteur des traits est construit comme représentation de la phrase ; ce vecteur est considéré comme un exemple positif si les deux concepts sont en relation dans BabelNet, comme un exemple négatif sinon.

- **Etape 2** : Apprentissage



La phase d'apprentissage consiste à entraîner un modèle statistique basé sur l'algorithme de régression linéaire MaxEnt. Ce modèle réalise une classification des entrées qui lui sont fournies sous forme de vecteurs. L'ensemble des exemples (positifs et négatifs) est découpé en deux parties, l'une pour entraîner le modèle, l'autre pour le tester. Une fois entraîné, le modèle est utilisé sur le jeu de test. Les résultats obtenus sont quantifiés afin de calculer les paramètres d'évaluation (précision, rappel et F-mesure).

- **Etape 3** : Extraction des relations



Ce modèle peut ensuite être utilisé sur tout corpus pour en extraire des relations en suivant le scénario suivant :

— Le corpus doit être traité comme lors de l'étape de construction des exemples, à l'aide des logiciels présents dans GATE.
— l'extracteur de termes est utilisé pour repérer des termes dans chaque phrase du corpus. Deux termes par phrase sont sélectionnés pour être des candidats à former une relation.
— un vecteur de traits est défini pour représenter chaque phrase, selon le même procédé que pour constituer les traits en phase de construction des exemples.
— ces vecteurs de trait sont fournis en entrée du modèle d'apprentissage, qui fournit en sortie une classification de ces traits. Les traits classés comme hyperonymes permettent de retrouver les phrases et les paires de termes en relation.

4.3 Choix des traits et repérage des termes

Pour chaque exemple, nous avons dû définir l'ensemble des propriétés exploitées par le système d'apprentissage, et la manière d'identifier les termes.

4.3.1 Choix des propriétés caractérisant les exemples

Pour représenter un exemple selon le format attendu par l'algorithme d'apprentissage, à savoir un vecteur de valeurs (numériques ou booléennes). Pour cela, nous devons sélectionner des propriétés permettant de caractériser une phrase, et décider de représenter chacune d'elle dans une des coordonnées du vecteur. Nous avons retenu les propriétés suivantes :

1. une paire de termes (ci-après Terme1 et Terme2) extraits de la phrase,
2. un contexte (ou fenêtre) de taille n formé par la séquence {n tokens précédant Terme1, Terme1, tokens séparant Terme1 et Terme2, Terme2, n tokens suivant Terme2},
3. un ensemble de valeurs de traits ayant différents niveaux de granularité (voir Table 4.1), et
4. la classe (positif ou négatif) à laquelle appartient l'exemple. Un exemple est positif (resp. négatif) si les deux termes dénotent deux concepts qui existent dans la ressource sémantique, et si la relation d'hyperonymie entre ces deux concepts est représentée (resp. n'est pas présente) dans cette ressource. Dans tous les autres cas, la paire de termes ne constitue pas un exemple d'apprentissage.

4.3.2 Choix des traits

Nous avons expérimenté deux variantes selon les traits retenus afin de juger de leur pertinence et mesurer leur contribution à la qualité du modèle d'apprentissage.

1. Un premier niveau de traits consiste à ne prendre en compte que le lemme et la catégorie grammaticale des mots.

2. En nous appuyant sur les travaux antérieurs de l'état de l'art, nous avons enrichi ces traits en prenant en compte deux informations : des mesures de distance entre les mots, et la présence ou l'absence d'une forme verbale entre les mots. L'ensemble est présenté dans la table 4.1.

Niveau de granularité	Trait	Signification	Type
Token	POS	Part Of Speech	chaîne de caractères
	LEMME	Forme lemmatisée du token	chaîne de caractères
Fenêtre	distT1	Nombre de tokens entre le mot et Terme1	entier
	distT2	Nombre de mots entre le token et Terme2	entier
	distT1T2	Nombre de tokens entre Terme1 et Terme2	entier
	nbMotsFenêtre	Nombre de tokens dans la fenêtre	entier
Phrase	nbMotsPhrase	Nombre de tokens dans la phrase	entier
	presVerbe	Présence d'une forme verbale	booléen

Table 4.1 – Ensemble des traits associés à un exemple.

4.3.3 Identification des termes

Nous avons successivement envisagé deux approches pour identifier les termes présents dans une phrase, et pour lesquels une relation va être recherchée : l'utilisation d'un extracteur de termes (en l'occurrence l'extracteur Yatea qui fonctionne sur le français) ; l'utilisation de termes de BabelNet.

Nous avons pour cela constitué une ressource, que nous appelons LBabel, qui rassemble tous les mots et syntagmes en langue française mentionnés dans les concepts (appelés synsets) de BabelNet. La constitution de cette liste est non triviale, dans la mesure où l'API de BabelNet permettant de faire des requêtes sur la ressource ne fournissait les termes en entier mais seulement les têtes des syntagmes. Le problème a été signalé et pris en compte dans notre algorithme. Grâce à cette ressource, nous avons pu réduire le temps de réalisation des expérimentations en évitant d'interroger systématiquement BabelNet. Elle a permis aussi d'éviter le recours à un extracteur de terme dans les phrases.

Ainsi, au lieu de rechercher dans BabelNet les termes T1 et T2 identifiés dans une phrase à l'aide de Yatea, nous avons projeté les termes de LBabel sur les phrases du corpus. Cela évite de lancer une exécution de Yatea, qui peut être coûteuse en temps. De plus, Yatea extrait des termes qui ne sont pas nécessairement présents dans BabelNet. Or nous cherchons ensuite ces termes dans BabelNet pour savoir s'ils sont en relation. Autant donc prendre directement comme termes potentiellement en relation des termes BabelNet présents dans les phrases.

Identifier des termes dans une phrase revient alors à projeter des termes BabelNet sur cette phrase, et donc de retenir une distance mesurant la similarité entre termes. Pour cela, nous avons utilisé une distance d'édition classique de type distance de Levenchtein. Une autre question est de savoir si, parmi les différents termes BabelNet reconnus pour un token de la phrase, on retient le plus précis (et donc le plus long) ou le plus général (et donc le plus court, qui est la tête des termes composés) ou si on les garde tous. Le choix a été fait de prendre potentiellement tous les termes possibles, puis de choisir aléatoirement une paire dont on recherche si elles est en relation dans BabelNet.

4.3.4 Illustration

Nous décrivons ci-dessous la construction d'un exemple à partir de la phrase *"Lime ou citron vert, le fruit des limettiers : Citrus aurantiifolia et Citrus latifolia"*

La projection de la liste des termes LBabel conduit à annoter la phrase par le termes `Lime`, `citron`, `citron vert`, `vert`, `fruit`. Considérons le couple <`Lime`, `fruit`> choisi aléatoirement par le système : Terme1=`Lime` et Terme2=`fruit`. Pour une fenêtre égale à 3 tokens [1], le système considère alors la phrase ainsi :

1. Nous avons évalué des fenêtres de dimension 1, 3 et 5, l'optimum étant obtenu pour la dimension 3

```
Terme1 ou citron vert, le Terme2 des limettiers :
```
où les mots correspondant aux termes ont été remplacés par Terme1 et Terme2. L'annotation par Tree-Tagger permet de remplacer les formes exactes des tokens par leur lemme précédé de leur catégorie syntaxique. La phrase devient :
```
 Terme1 KON/ou NOM/citron ADJ/vert PUN/, DET:ART/le Terme2
PRP:det/du NOM/limettier PUN/:
```

Enfin, des fonctions de traits sont calculées pour définir les valeurs de ces traits. Elles associent à chaque token de la phrase les distances relatives (en nombre de tokens) de ce token à Terme1 et à Terme2 sous forme de couple de valeurs, le nombre de tokens entre Terme1 et Terme2 (en l'occurrence 5) et le nombre de tokens dans la phrase (ici 16). Le dernier trait indique l'absence de verbe dans la phrase.
```
(0,6) (-1,5) (-2,4) (-3,3) (-4,2) (-5,1) (-6,0) (-7,-1) (-8,-2) (-9,-3) 5 16
false
```

Voici l'exemple dans son intégralité :

```
 Terme1 KON/ou NOM/citron ADJ/vert PUN/, DET:ART/le Terme2
 PRP:det/du NOM/limettier PUN/:
 (0,6) (-1,5) (-2,4) (-3,3) (-4,2) (-5,1) (-6,0) (-7,-1) (-8,-2) (-9,-3) 5 16
false
```

Cet exemple est positif car les termes "lime" et "fruit" renvoient à des ressources en relation d'hyperonymie dans BabelNet.

4.4 Mise en œuvre et résultats

Nous avons produit automatiquement ~8000 exemples parmi lesquels avons conservé 6000 exemples (3000 positifs et 3000 négatifs). L'ensemble d'entraînement est composé de 4000 exemples pris aléatoirement parmi les 6000, en maintenant une quasi-parité positifs / négatifs (~2000/~2000), et l'ensemble de test comporte les 2000 exemples restants. Nous avons entraîné un algorithme de régression logistique binaire, MaxEnt [Berger et al., 1996] sur l'ensemble d'entraînement.

4.4.1 Application aux pages de désambiguïsation

Cette expérience s'est appuyée sur le corpus de référence annoté décrit dans le chapitre précédent. Du fait de la projection automatique des labels de BabelNet sur le corpus, nous n'avons plus recours à un extracteur de termes, ce qui simplifie la chaîne de traitement décrite en section 4.2.

L'évaluation est réalisée à deux niveaux :

— Evaluation de la performance du classifieur
— Evaluation de l'apport de l'extraction pour l'enrichissement de la base de connaissances

Nous avons en tout réalisé 4 expérimentations en modifiant à chaque fois les vecteurs de traits des exemples pour choisir la meilleure composition. A chaque fois, le modèle d'apprentissage a été évalué en sur un corpus de 20 pages, en confrontant les relations d'hyperonymie trouvées par le modèles aux relations annotées à la main par 2 annotateurs. Le nombre total des relations annotées est de 688. Les résultats en terme de rappel et précision des 4 expérimentations réalisées sont présentés dans le tableau 4.2.

Expé.	Traits utilisés	Rappel	Précision
1	3 mots + POS + LEMME + distT1T2	0.46	0.81
2	5 mots + POS + LEMME + distT1T2	0.48	0.67
3	1 mot + POS + LEMME + distT1T2	0.47	0.65
4	3 mots + POS + LEMME + distT1T2 +distT1 + distT2 + nbMotsPhrase + presVerbe	0.58	0.72

TABLE 4.2 – Expérimentations réalisées sur les pages d'homonymie

Nous remarquons que les paramètres de l'expérimentation 4 ont permis d'obtenir les meilleurs résultats en terme de rappel et précision. C'est donc cette configuration que nous avons validée pour étudier la complémentarité de cette approche avec l'approche par patrons, dans le chapitre suivant.

4.4.2 Application à l'ensemble du corpus Wikipedia

Annotation manuelle

L'évaluation de la méthode sur l'ensemble du corpus Wijipédia a nécessité de mettre en œuvre une nouvelle tâche d'annotation manuelle, puisque nous ne disposions que de l'annotation de pages Wikipedia issues de l'ensemble spécifique des pages de désambiguïsation. Cette nouvelle annotation a été réalisée par 2 annotateurs. Elle a été menée sur les premiers paragraphes de 56 pages Wikipedia extraites aléatoirement du corpus, soit 300 paires de segments en relation hyperonymique. L'annotation a été menée sur le corpus brut en amont de l'étape d'extraction des termes. Elle a donc consisté à délimiter les segments et à poser la relation.

A titre d'exemple :

Parmi les autres attributs figurent la queue de taureau fixée à l'arrière du pagne, la barbe cérémonielle, les sandales et l'étui-mekes.
→ hyperonyme(*autres attributs*,*barbe cérémonielle*)

L'annotation proprement dite a été précédée d'une phase d'annotation test et de mise en commun, pour repérer les décalages et établir les règles précises d'annotation, en particulier concernant la délimitation des segments. A l'issue de l'annotation séparée des 300 paires, une phase d'adjudication a permis la production d'un fichier unique. Les principes d'annotation mis en œuvre sont les suivants :

- **Taille du contexte :** on ne considère une relation qu'à l'intérieur d'une phrase (et non à cheval sur plusieurs phrases) et on peut valider plusieurs relations dans la même phrase ;
- **Délimitation des segments :**
  - On ne reconstitue aucun segment implicite ;
  - On ne réalise pas de lemmatisation, la forme de surface est donc conservée ;
  - On supprime le déterminant, mais on conserve les modifieurs qui précèdent le terme ;
  - On ne prend pas toujours le terme maximal, on supprime les modifieurs qui suivent dans les cas suivants : les relatives, les participes ;
- **Relation d'hyperonymie :**
  - On ne traite pas les cas d'inclusion lexicale qui amènent à décomposer le terme ;
  - On ne considère pas qu'il y a une relation hyperonymique dans les cas d'équivalence, de synonymie ;
  - En cas de doute, on vérifie par le test « est un type de » la validité du terme hyperonmyique (est-il un bon catégorisateur).

Dans cette expérience, ce n'est pas LBabel mais un sous-ensemble de celle-ci, à savoir la liste des termes BabelNet présents dans le fichier construit par les annotateurs, qui servira ensuite pour l'évaluation de l'expérimentation. C'est donc cette liste qui sera projetée sur le corpus et on ne cherchera des relations qu'entre ces termes. Il s'agira donc de prendre une décision pour l'appariement : soit rechercher toutes les décompositions, soit ne considérer que la tête nominale du terme. C'est la deuxième option, plus simple, qui a été choisie.

Prise en compte de traits supplémentaires

Les expérimentations menées sur le corpus Wikipedia intégral ont intégré des traits supplémentaires par rapport aux expériences menées sur le corpus d'homonymie.

Information distributionnelle Comme nous l'avons vu dans le chapitre consacré à l'état de l'art, l'information distributionnelle peut être utilisée comme un indice complémentaire pour renforcer l'hypothèse d'une relation d'hyperonymie : en effet, les termes hyponymes et hyperonymes ont tendance à partager les mêmes contextes, même si une dissymétrie existe : l'hyperonyme, parce qu'il présente des caractéristiques conceptuelles plus génériques, peut apparaître dans un ensemble plus large de contextes que l'hyponyme [Lenci and Benotto, 2012b, Kotlerman et al., 2010]. Si l'on sait que ces indices ne permettent pas, à eux seuls, de discriminer la relation d'hyperonymie, leur intégration dans le processus d'apprentissage parmi d'autres traits est une piste intéressante, que nous avons voulu tester. Pour cela, nous avons intégré dans les traits d'apprentissage une information relative à la

mesure de similarité distributionnelle (valeur du cosinus) des deux termes candidats à la relation d'hyperonymie. Cette valeur est calculée à partir du modèle distributionnel produit par Word2Vec (paramètres par défaut) sur le corpus Wikipedia. Elle est intégrée dans l'apprentissage sous la forme d'un trait binaire : si les deux termes présentent une valeur de similarité supérieure ou égale à un seuil donné, la valeur est "oui", elle est "non" dans le cas contraire. Nous avons testé plusieurs valeurs de seuil (0.3, 0.4, 0.5). D'autres façons d'intégrer l'information distributionnelle ont été également testées : intégration de la valeur de similarité, intégration des vecteurs de traits des 2 termes. Seule l'expérience recourant à une valeur binaire et à un seuil de similarité est reproduite ici.

Prise en compte de la présence d'un patron Nous avons envisagé une autre manière d'intégrer l'information linguistique liée à la présence d'un patron d'hyperonymie dans le contexte des termes. Elle consiste à ajouter un trait qui teste la présence d'un des patrons recensés dans la ressource fusionnée CAMELEON + MAR-REL.

Résultats

Nous reproduisons dans le tableau 4.3 les résultats obtenus par l'application successive des traits que nous venons de présenter.

Expé.	Précision	Rappel	F-mesure
Traits élémentaires	0.71	0.65	0.68
Traits enrichis	0.73	0.65	0.69
Traits enrichis + valeur distrib. seuil >0.3	0.74	0.65	0.69
Traits enrichis + valeur distrib. seuil >0.4	0.74	0.65	0.69
Traits enrichis + valeur distrib. seuil >0.5	0.75	0.65	0.7
Traits enrichis + valeur distrib. >0.5 + test présence patron	0.79	0.72	0.75

TABLE 4.3 – Expérimentations réalisées sur l'ensemble du corpus Wikipedia

Rappelons que les traits élémentaires correspondent aux seules informations relatives au POS et au lemme. Les traits enrichis incorporent l'ensemble des informations décrites dans la section précédente, soit : nombre de tokens entre T1 et T2 + présence d'un verbe + taille de la phrase.

4.4.3 Bilan de l'expérimentation

On constate que l'enrichissement des traits a un impact positif très limité sur les performances. La combinaison de traits la plus pertinente est celle qui intègre la présence d'un patron, mais il faudrait tester les traits indépendamment pour mesurer leur impact propre, ce qui reste à faire. L'apport des traits distributionnels est également négligeable, ce qui est surprenant par rapport à l'état de l'art où la prise en compte de la sémantique via des vecteurs de mots conduit souvent à une amélioration significative. Provisoirement, ces résultats nous ont amenés à renoncer à ces traits dans l'étape ultérieure de l'expérimentation, qui a consisté à tester la combinaison de l'approche par patrons et de l'approche par apprentissage. A plus long terme, nous envisageons de prendre en compte les vecteurs des mots T1 et T2 de manière différente.

4.5 Conclusion

La mise au point d'une démarche d'apprentissage automatique pour l'extraction de relations sémantiques est une tâche complexe. L'apprentissage automatique étant avant tout un processus de classification statistique, il convient de définir :

— le corpus d'étude : nous avons étudié deux corpus aux caractéristiques différentes : les pages de désambiguïsation de Wikipedia, dont le texte n'est pas rédigé et où la ponctuation et la mise en forme jouent un rôle important pour exprimer les relations, et les paragraphes de définition de Wikipédia qui, eux, contiennent du texte rédigé ;

— ce qu'on entend par terme dans une phrase et par paire de termes dont on doit vérifier s'ils sont en relation : nous avons d'abord considéré les termes identifiés par un extracteur de termes pour le français, puis les termes français présents dans la ressource sémantique BabelNet ;

— les éléments à classer : il s'agit des phrases contenant des couples de termes dont on veut identifier s'ils sont en relation d'hyperonymie ou non ;
— une représentation adaptée des informations à classer (donc de ces phrases), à savoir des traits à retenir pour former des vecteurs de traits : nous avons choisi des traits classiques et évalué l'apport de traits basés sur des vecteurs distributionnels ; nous avons testé plusieurs valeurs de seuil pour comparer les vecteurs, et plusieurs ensembles de traits ;
— des jeux d'exemples d'entraînement, composés d'exemples positifs et négatifs : pour cela, nous avons choisi une approche par supervision distante, de manière à générer automatiquement des exemples en utilisant les couples de termes en relation dans la ressource sémantique BabelNet ;
— un jeu de données d'évaluation, afin d'évaluer les différentes approches testées ; nous avons annoté manuellement dans chacun des cas plusieurs pages de chaque corpus de manière à obtenir un nombre significatif d'exemples (300 dans chaque cas).

Le meilleur des résultats ont été obtenus avec des traits classiques, une fenêtre de 3 mots et des termes tirés de BabelNet. La F-mesure est alors de 70% contre 68% avec les traits élémentaires, ce qui représente un gain peu significatif. On voit sur le tableau 4.3 que la dernière ligne de résultats améliore plus significativement la F-mesure. Elle correspond à une nouvelle expérience qui étend celles-ci par une comparaison de l'approche par apprentissage à l'approche par patron présentée au chapitre 3. Nous développons cette nouvelle étude dans le chapitre 5.

5 Combinaison de méthodes pour l'extraction de relations d'hyperonymie

Sommaire

5.1 Introduction

Après la présentation de chacune des deux approches (linguistique vs par apprentissage) que nous avons mises en œuvre respectivement dans les chapitres 3 et 4, ce chapitre décrit les expérimentations visant à tester l'intérêt d'appliquer de manière complémentaire différentes approches sur un même corpus pour identifier des occurrences de la relation d'hyperonymie, à travers ses différents modes d'expression.

Pour ce faire, nous avons appliqué ces deux approches sur le même corpus. Nous avons choisi le corpus des pages de désambiguïsation de Wikipedia. En effet, commenous l'avons souligné au chapitre 3, ces pages offrent un premier cas de figure favorable pour appliquer des patrons et de l'apprentissage : très riches en relations d'hyperonymie, elles comportent du texte rédigé (assez minoritairement), et, pour l'essentiel, du texte peu rédigé (structure syntaxique incomplète) usant de mise en forme matérielle variée comme la ponctuation, diverses polices de caractère ou la disposition.

L'intérêt de cette étude est de tirer profit des avantages de chacune des deux approches, notamment d'associer la bonne précision de l'approche par patrons et le bon rappel de l'approche par supervision distante. De plus, nous avons mené une étude qualitative pour évaluer l'apport de la combinaison de ces deux approches : quelles sont les relations qui sont efficacement trouvées par les patrons et ne sont pas trouvées par l'approche statistique, et inversement ? Cette évaluation qualitative permettrait de cerner les limites de chaque approche.

Ce chapitre présente cette étude sur la complémentarité des approches, qui a fait l'objet de 3 articles de conférence. Nous présentons d'abord la méthode mise en place pour évaluer cette complémentarité et fournissons ensuite 2 des trois articles qui en présentent les résultats.

5.2 Présentation de la méthode

Les grandes étapes de la méthode sont assez classiques : il s'agit de mettre en oeuvre les deux méthodes avec les meilleurs paramètres pour chacune, et de définir une 3e méthode, dans laquelle la reconnaissance d'un patron dans la phrase est utilisée comme un trait complémentaire dans la représentation d'un exemple. Ensuite, les méthodes sont comparées toujours selon les critères de précision, rappet et F-mesure. Nous avons donc 4 méthodes comparées :

— l'approche par patron

Chaque corpus est étiqueté morpho-syntaxiquement par TreeTagger [1]. Pour identifier l'expression de relations sémantiques, le texte est aussi annoté à l'aide de termes, à savoir des syntagmes, le plus souvent nominaux, pouvant désigner des entités ou des classes conceptuelles. Les termes peuvent donc être inclus les uns dans les autres (*système* dans *système solaire*). Plutôt que d'utiliser un extracteur de termes, nous avons choisi de construire a priori deux listes de termes :

— LBabel comporte les labels en français des concepts présents dans la ressource sémantique BabelNet [2] ; elle servira à annoter le corpus d'apprentissage ;

1. http ://www.cis.uni-muenchen.de/schmid/tools/TreeTagger
2. http ://babelnet.org/

— LCorpus, composée des termes annotés manuellement dans le corpus de référence ; elle sera utilisée pour l'évaluation de l'approche sur le corpus de référence.

L'annotation du corpus d'apprentissage par des termes provenant d'une source sémantique partagée assure une plus grande validité du modèle d'apprentissage. Cela évite aussi que l'identification des termes vienne biaiser l'extraction des relations.

5.3 Résultats et Évaluation

A partir du corpus de référence, 688 exemples vrais positifs (VP) et 267 exemples vrais négatifs (VN) ont été identifiés manuellement. Nous interprétons les résultats obtenus afin de juger de la complémentarité des méthodes expérimentées, et de l'intérêt de leur utilisation conjointe.

5.3.1 Évaluation quantitative

Nous avons mis en œuvre les deux approches indépendamment, l'approche par patrons avec **PatronsG** puis avec **PatronsGS**, et l'approche par apprentissage supervisé (MaxEnt). Nous avons évalué leur complémentarité à l'aide de l'union et de l'intersection de leurs résultats. Le tableau 5.1 fournit les différentes valeurs de la précision, du rappel, de la F-mesure et de l'exactitude.

	PatronsG	PatronGS	MaxEnt	PatronsG union MaxEnt	PatronsGS union MaxEnt
Précision	0.96	0.81	0.71	0.72	0.73
Rappel	0.04	0.46	0.63	0.65	0.77
F-mesure	0.07	0.59	0.67	0.68	0.75
Exactitude	0.31	0.54	0.55	0.56	0.63

TABLE 5.1 – Évaluation des approches.

Confirmant l'état de l'art Hearst [1992] Malaisé et al. [2004], les patrons obtiennent un fort taux de précision, au détriment d'un faible rappel. Et comme attendu, les patrons génériques augmentés des patrons spécifiques donnent de meilleurs résultats que les patrons génériques seuls : bien que la précision baisse à 0.81, le rappel passe à 0.46, pour une F-mesure de 0.59. En revanche, l'approche par apprentissage supervisé, moins bonne en précision, produit un meilleur taux de rappel. En effet, le corpus présente de fortes régularités (aussi bien syntaxiques que de mise en forme), ce qui permet de renforcer l'apprentissage. Ces résultats corroborent les résultats reportés dans la littérature pour d'autres types de corpus.

Nous constatons que l'application des deux approches fournit une meilleure F-mesure que les approches prises indépendamment, et que l'union des résultats de PatronsGS et de MaxEnt permet d'obtenir la meilleure F-mesure. Nous avons également pu analyser la complémentarité des PatronsGS et de MaxEnt à travers les résultats donnés dans la Table 5.2.

	PatronsGS ou MaxEnt	PatronsGS et MaxEnt	PatronsGS seul	MaxEnt seul	Aucune des 2 méthodes
Nombre VP	527	221	96	210	161

TABLE 5.2 – Parmi les 688 VP du corpus de référence, nombre de VP trouvés par les approches.

Nous avons pu ainsi constater que parmi les 306 VP qui ne sont identifiés que par une seule des deux méthodes, PatronsGS en identifie 32%, et MaxEnt 68%. Ce résultat confirme la complémentarité de ces deux approches.

5.3.2 Evaluation qualitative

Nous avons constaté que parmi les 221 relations trouvées par PatronsGS et MaxEnt, 9 relations concernent des relations exprimées par le verbe *être*, comme entre les termes *macédoine* et *salade*

de fruits dans la phrase "La macédoine est une salade de fruits ou de légumes". Quasiment toutes les autres relations correspondent au schéma "X, Y" comme dans "Le cheval de Troie, un mythe grec". On remarque toutefois que les patrons retrouvent des relations entre des noms communs, alors que MaxEnt trouve essentiellement des relations entre entités.

Pour les 96 relations trouvées par PatronsGS et n'ayant pas été identifiées par MaxEnt, on trouve des relations (19) exprimées par le verbe *être*, mais lorsque la relation n'est pas exprimée en début de phrase, comme la relation entre *Babel fish* et *espèce imaginaire* dans "Le poisson Babel ou Babel fish est une espèce imaginaire". Quasiment toutes les autres relations trouvées seulement par les patrons correspondent au schéma "X,Y" comme décrit ci-dessus. Nous n'avons pas encore identifié la cause du silence de MaxEnt à ce niveau.

Parmi les 210 relations trouvées par MaxEnt et non identifiées par PatronsGS, on trouve les cas d'inclusion lexicale (85), comme la relation entre *gare de Paris Bastille* et *gare*, issue du groupe nominal "gare de Paris Bastille". MaxEnt permet également de trouver des relations exprimées par d'autres verbes d'état (8) comme la relation entre *aigle* et *oiseaux rapaces* dans "Aigle désigne en français certains grands oiseaux rapaces". MaxEnt identifie aussi des relations exprimées dans des unités textuelles comportant des coordinations, comme la relation entre *poisson Babel* et *espèce imaginaire* dans "Le poisson Babel ou Babel fish est une espèce imaginaire", ou encore entre *Louis Babel* et *explorateur* dans "Louis Babel, prêtre-missionnaire oblat et explorateur". Enfin MaxEnt identifie très bien les relations au sein de texte usant de mise en forme comme la relation entre *arête* et *barbe de l'épi* dans "Arête, "barbe de l'épi"", ou entre *Aigle* et *chasseur de mines* dans "Aigle (M647), chasseur de mines".

Finalement, parmi les 161 relations qui n'ont été trouvées ni par PatronsGS ni par MaxEnt, 55 correspondent à des inclusions lexicales (nous n'avons pas non plus identifié la cause du silence de MaxEnt), 64 possèdent une incise entre les deux termes comme dans "Un Appelant (jansénisme) est, au XVIIIe siècle, un ecclésiastique qui ... " ou "Giuseppe Cesare Abba (1838-1910), écrivain". Les 42 cas restants concernent des formes d'expression non prises en charge par les patrons et trop peu fréquentes pour être apprises par MaxEnt, comme "X tel que Y".

Cette analyse confirme l'intérêt de mettre en œuvre sur un même corpus des approches complémentaires. Tout d'abord, nous avons pu constater que les inclusions lexicales sont identifiées par MaxEnt seul. Ensuite, les différentes occurrences de relations au sein d'une même phrase sont identifiées par les deux méthodes, comme on l'a vu ci-dessus à travers l'exemple "Le poisson Babel ou Babel fish est une espèce imaginaire". Enfin, MaxEnt permet d'identifier les relations exprimées selon différentes variantes d'un même schéma (incises, usage de mise en forme, etc.), dès lors que ces structures sont récurrentes. Par ailleurs, nous avons également pu observer que PatronsGS et MaxEnt sont complémentaires dans une proportion de ~1/3 vs. 2/3.

5.4 Les articles tirés de cette étude

Trois articles ont été tirés de ce travail :

- Ghamnia, A., Kamel, M., Trojahn, C., Fabre, C., Aussenac-Gilles, N. (2017). Extraction de relations : combiner les techniques pour s'adapter à la diversité du texte. In *Actes des 28es Journées francophones d'Ingénierie des Connaissances IC 2017*,Caen, France. 86-97.
- Kamel, A., Trojahn, C., Ghamnia, A., Aussenac-Gilles, N., Fabre, C. (2017). Extracting hypernym relations from Wikipedia disambiguation pages : comparing symbolic and machine learning approaches. In *Proceedings of the International Conference on Computational Semantics (IWCS)*, Montpellier, France.
- Kamel, A., Trojahn, C., Ghamnia, A., Aussenac-Gilles, N., Fabre, C. (2017). A Distant Learning Approach for Extracting Hypernym Relations from Wikipedia Disambiguation Pages. In *Proceedings of the 21st International Conference on Knowledge-Based and Intelligent Information Engineering Systems (KES 2017)*, Marseille, France. 1764-1773.

Nous reproduisons les 2 derniers dans les pages suivantes.

Extracting hypernym relations from Wikipedia disambiguation pages: comparing symbolic and machine learning approaches

Mouna Kamel[1], Cassia Trojahn[1], Adel Ghamnia[1,2], Nathalie Aussenac-Gilles[1], Cécile Fabre[2]
[1] Institut de Recherche en Informatique de Toulouse, Toulouse, France
{mouna.kamel,cassia.trojahn,adel.ghamnia,nathalie.aussenac-gilles}@irit.fr
[2] Laboratoire CLLE, équipe ERSS, Toulouse, France
cecile.fabre@univ-tlse2.fr

Abstract

Extracting hypernym relations from text is one of the key steps in the construction and enrichment of semantic resources. Several methods have been exploited in a variety of propositions in the literature. However, the strengths of each approach on a same corpus are still poorly identified in order to better take advantage of their complementarity. In this paper, we study how complementary two approaches of different nature are when identifying hypernym relations on a structured corpus containing both well-written text and syntactically poor formulations, together with a rich formatting. A symbolic approach based on lexico-syntactic patterns and a statistical approach using a supervised learning method are applied to a sub-corpus of Wikipedia in French, composed of disambiguation pages. These pages, particularly rich in hypernym relations, contain both kinks of formulations. We compared the results of each approach independently of each other and compared the performance when combining together their individual results. We obtain the best results in the latter case, with an F-measure of 0.75. In addition, 55% of the identified relations, with respect to a reference corpus, are not expressed in the French DBPedia and could be used to enrich this resource.

1 Introduction

In many fields such as artificial intelligence, semantic web, software engineering or information retrieval, applications require a strong reasoning ability, based on semantic resources that describe concepts and the relations between them. These resources can be manually designed. They are of good quality, however due to the high cost of their design, they offer a limited domain coverage. With the increasing amount of textual documents available in digital format, NLP processing chains offer a good support to design such resources from text. In this context, the task of automatically extracting relations from text is a crucial step (Buitelaar et al., 2005). Numerous studies have attempted to extract hypernym relations, as they allow for expressing the backbone structure of such resources and for assigning types to entities.

While symbolic approaches usually rely on manually defined lexico-syntactic patterns identifying clues of relations between terms (Hearst, 1992), statistical approaches, which are nowadays predominant, are generally based on supervised (Pantel and Pennacchiotti, 2008) or unsupervised (Banko et al., 2007) learning methods, or on distributional spaces (Lenci and Benotto, 2012). These methods of different nature answer to the need of exploiting corpora with different specificities (e.g. domain granularity, nature of the corpus, language, target semantic resource, etc.) and which express the hypernym relation in different forms. For giving some examples, this kind of relation can be expressed by the lexicon and the syntactic structure as in the sentence *sand is a sedimentary rock*, by a lexical inclusion as in *domestic pigeon* (implied *domestic pigeon is a pigeon*), or by using punctuation or layout features that replace lexical markers like the comma in *Trojan horse, a Greek myth* or even the disposition in enumerative structures.

The study we conduct in this paper aims to show the interest of applying several approaches on a same corpus in order to identify hypernym relations through their various forms of expression. We are

particularly interested in exploiting a corpus containing both well-written text (i.e., sentences expressed with a complete syntactic structure) and syntactically poor formulations (i.e., sentences with syntactic holes), together with a rich formatting. We analyze the complementarity of a symbolic approach based on lexico-syntactic patterns and a statistical approach based on supervised learning. We applied these two approaches to a corpus of Wikipedia disambiguation pages, which are very rich in hypernym relations differently expressed, as these pages contain both well-written text and poorly-written text.

Our proposal focuses on the combination of the individual results rather than on the combination of the approaches themselves (e.g., by learning patterns). Indeed, combining patterns with machine learning usually relies on path-based methods (Snow et al., 2004) (Snow et al., 2006) (Riedel et al., 2013). However, dependency parsers have proven to perform worse on poorly-written text. Although our approach is naive in that sense, it proves to provide good results, in particular, in terms of F-measure.

This work is part of the SemPedia[1] project that aims at enriching the semantic resource DBPedia for French (semantic resources targeting this language are scarce), by proposing a new Wikipedia extractors dedicated to the hypernym relation. Hence, we evaluate how the extracted relations could potentially enrich such kind of resource.

The paper is organized as follows. Section 2 outlines the main work related to our proposal. Section 3 presents the materials and methods used in our study, namely the description of the training and reference corpus, their pre-processing, and the extraction approaches. The results obtained are presented and discussed in Section 4. Finally, Section 5 concludes the paper and presents future directions.

2 Related work

In the field of relation extraction, the pioneering work of the symbolic methods is that of Hearst (Hearst, 1992) which defined a set of lexico-syntactic patterns specific to the hypernym relation for English. This work has been adapted and extended to French for the hypernym relation (Morin and Jacquemin, 2004), for the meronymic relation (Berland and Charniak, 1999), and for different types of relations (Séguéla and Aussenac-Gilles, 1999), by progressively integrating learning techniques.

With respect to statistical approaches and, in particular, those based on machine learning, which are specially required when dealing with large corpus, Snow and colleagues (Snow et al., 2004) and Bunescu and Mooney (Bunescu and Mooney, 2005) apply supervised learning methods to a set of manually annotated examples. While the cost of manual annotation is the main limitation of supervised learning, distant supervision method consists in building the set of examples using an external resource to automatically annotate the learning examples (Mintz et al., 2009). Another way to avoid manual annotation is the semi-supervised learning method called bootstrapping which uses a selection of patterns to construct the set of examples (Brin, 1998). Agichtein and Gravano (Agichtein and Gravano, 2000), and Etzioni and colleagues (Etzioni et al., 2004) have used this method by adding semantic features to identify relations between named entities. Unsupervised learning, based on clustering techniques, was implemented by Yates and colleagues (Yates et al., 2007) and Fader and colleagues (Fader et al., 2011) which used syntactic features to train their classifiers relations between named entities. Some of these works are also based on distributional analyses (Kotlerman et al., 2010) (Lenci and Benotto, 2012) (Fabre et al., 2014). In the work of Kotlerman and colleagues (Kotlerman et al., 2010), they quantify distributional feature inclusion, where the contexts of a hyponym are expected to be largely included in those of its hypernym. Lenci and Benotto (Lenci and Benotto, 2012) explore the possibility of identifying hypernyms using a directional similarity measure that takes into account not only the inclusion of the features of u in v, but also the non-inclusion of the features v in u. The hypothesis from Santus and colleagues (Santus et al., 2014) is that most typical linguistic contexts of a hypernym are less informative than those of its hyponyms.

Beyond these works, which evaluate approaches independently of each other, few results have been reported on the respective contributions and the complementarity of methods. Granada (Granada, 2015) compared the performance of different methods (patterns-based, head-modifier, and distributional ones)

[1] http://www.irit.fr/Sempedia

for the task of hypernym relation extraction in different languages, by defining several metrics such that density and depth of hierarchies. The evaluation was carried out on different types of corpus but does not take into account the learning approaches. Yap and Baldwin (Yap and Baldwin, 2009) study the impact of the corpus and the size of training sets on the performance of similar supervised methods, on the extraction of several types of relation (hypernym, synonymy and antonymy), whereas Abacha and Zweigenbaum (Ben Abacha and Zweigenbaum, 2011) combine patterns and a statistical learning method based on the SVM classifier for extracting relations between specific entities (disease and treatment) from a biomedical corpus. In the same line, we exploit methods of different nature, focusing on the specific hypernym relation.

In particular, with respect to the approaches combining patterns and distributional methods, most of them rely on path-based methods. It is the case, for instance, of the learning approach of Snow and colleagues (Snow et al., 2004), which automatically learned pattern spaces based on syntactic dependency paths. These paths represent the relationship between hypernym/hyponym word pairs from WordNet and are used as features in a logistic regression classifier. Variations of this method have been applied in different tasks, such as hypernym extraction (Snow et al., 2006) (Riedel et al., 2013) and extraction of definitions (Navigli and Velardi, 2010). However, as stated in (Kotlerman et al., 2010), one major limitation in relying on lexico-syntactic paths is the sparsity of the feature space, since similar paths may somewhat vary at the lexical level. In (Nakashole et al., 2012), generalizing such variations into more abstract paths proved to improve the results, in particular recall. On the other hand, while those approaches mostly rely on dependency trees extracted from well-written text, the performance of dependency parsers has proven to be very low on poorly-written corpus. In that sense, our focus here is to exploit strategies fitting a corpus rich in poorly-written text and where the polysemous occurs frequently (for instance, for the term "Didier Porte", "porte" is tagged as a verb instead of a noun). It is one of the reasons we focus here on the complementarity of the approaches rather than on their combination.

Finally, with respect to the enrichment of DBPedia knowledge base, several tools, called "extractors" have been developed to extract relations from the different elements present in the Wikipedia pages. Morsey and colleagues (Morsey et al., 2012) developed 19 extractors for analyzing abstract, images, infobox, etc. Other works focus on the hypernym relation. For example, Suchanek and colleagues (Suchanek et al., 2007) used the 'Category' part of Wikipedia pages to build the knowledge base, Yago, Kazama and Torisawa (Kazama and Torisawa, 2007) which exploited the 'Definition' part, and finally Sumida and Torisawa (Sumida and Torisawa, 2008) who were interested in the menu items. We can see that the DBPedia knowledge base is built essentially from the structural elements of the Wikipedia pages. Works targeting relation extraction from text have been exploited in a lesser extend (Rodriguez-Ferreira et al., 2016), which means that most of the knowledge in these pages remains under-exploited. Our aim here is to measure the degree of enrichment of semantic resources when exploiting this kind of relation extraction approach.

3 Material and methods

In this section, we describe the Wikipedia sub-corpus we used, its pre-processing, and the extraction methods we have considered.

3.1 Corpus

Different types of pages can be identified within the Wikipedia encyclopedia. Among them, the *disambiguation pages* list the articles whose title is polysemous, giving a definition of all the accepted meanings for this title, which refer to as many entities. Thanks to the Wikipedia's charter guidelines, which recommend the use of templates (for instance, *Toponyms*, *Patronyms*, etc.), these pages present editorial as well as formatting regularities for presenting the different meanings of the term on the page. For each meaning, a definition and a link to the corresponding page are provided. In fact, the definitions are textual objects in which the hypernym relation is often present (Malaisé et al., 2004) (Rebeyrolle

and Tanguy, 2000). Furthermore, on these pages, the definitions take varied but predictable forms. For instance, the following excerpt (Figure 1) which comes from the *Mercure* disambiguation page[2] shows different hypernym relations, expressed thanks to the lexicon (*le mercure est un élément chimique*), with the help of punctuations (the comma in *le Mercure, un fleuve du sud de l'Italie*), taking benefit from the lexical inclusion (*appareil de mesure*, implying that *appareil de mesure* is an *appareil*), or using dispositional and typographical characters (the structure substitutes the lack of complete syntax and expresses a good part of the text meaning) especially when expressing enumerative structure (*la diode à vapeur de mercure est un appareil de mesure*, *la pile au mercure est un appareil de mesure*, etc.).

Physique et chimie [modifier | modifier le code]

- Le mercure (symbole Hg) est un élément chimique.
- Le terme mercure rouge désignait au XIXe siècle l'iodure de mercure. Dans la dernière partie du XXe siècle il a été appliqué à une substance imaginaire, présentée comme un matériau stratégique rentrant dans la construction des armes nucléaires.
- Le millimètre de mercure (symbole mmHg), ou torr, est unité de mesure de pression.
- Plusieurs appareils de mesure ou méthodes physiques font référence au mercure, dont notamment :
  - la diode à vapeur de mercure,
  - la pile au mercure,
  - la pompe à mercure,
  - le porosimètre à mercure (en),
  - le thermomètre à mercure.

Toponyme et hydronyme [modifier | modifier le code]

Mercure est un nom de lieu notamment porté par :

- Mercure, une station du métro de Lille Métropole ;
- le Mercure, un fleuve du sud de l'Italie ;
- les îles Mercure, un archipel néo-zélandais, au large de la péninsule de Coromandel.
- le lac Mercure, un lac de l'île principale de l'archipel des Kerguelen, dans les Terres australes et antarctiques françaises ;
- le monastère Saint-Mercure, un important monastère féminin copte orthodoxe, situé dans le vieux Caire (Égypte) ;
- le mont Mercure, une montagne d'Italie ;
- Saint-Michel-Mont-Mercure, une ancienne commune française située dans le département de la Vendée, en région Pays-de-la-Loire
- la Vallée du Mercure, un grand bassin fluvial italien situé dans le sud de la Basilicate et le nord de la Calabre, et qui fut occupé par u lac au Pliocène.

Figure 1: Fragment of the disambiguation page *Mercure*.

We have compiled a corpus made of 5924 French disambiguation pages (XML version of the 2016 Wikipedia dump). From this corpus were extracted two sub-corpora:

- 20 randomly selected disambiguation pages form the *reference corpus*. In these pages, hypernymy relations were manually annotated, marking the terms referring to the related entities and the zone of the text where the relations were identified. This sub-corpus is used to qualitatively evaluate our approach and to evaluate the potential enrichment of DBPedia (Section 4.3);
- the remaining pages form the *training corpus*, which is intended to train and evaluate our learning model (Section 3.3.2).

3.2 Pre-processing

The content of each page has been labeled with morpho-syntactic tags, POS and lemma, using TreeTagger[3]. For identifying the expression of semantic relations, the text is also annotated using terms, namely syntagms, usually nominal, that may designate entities or conceptual classes. For example, *Mercure, système solaire, planète* (*Mercury, solar system, planet*) are some of the terms in Figure 1. The terms can therefore be included in each other (e.g., *system* in *solar system*). Rather than using a term extractor, we chose to construct *a priori* two lists of terms:

[2]https://fr.wikipedia.org/wiki/Mercure
[3]http://www.cis.uni-muenchen.de/schmid/tools/TreeTagger

- LBabel contains the list of terms retrieved from the French labels of concepts present in the semantic resource BabelNet[4]. This list will serve to train the learning system, as detailed in Section 3.3.2;
- LCorpus contains the list composed of the manually annotated terms from the reference corpus (Section 4).

These lists are then respectively projected on the pre-processed learning and reference corpora. In fact, the annotation of the corpus by terms derived from a shared semantic source ensures the validity of the learning model. This also prevents the identification of terms biasing the relation extraction process.

3.3 Relation extraction approaches

As already stated before, we have chosen two approaches of different nature which are often opposed by the cost of their implementation and by the precision and recall they provide: a symbolic approach based on lexical-syntactic patterns, and a statistical approach based on supervised learning using the distant supervision principle. While patterns represent recurring language patterns expressed through lexicon, syntactic and punctuation elements, automatic learning allows for combining features of different natures (morphological, syntactic, semantic or shaping) and for capturing the properties of contexts in a more global way. These approaches are detailed below.

3.3.1 Lexico-syntactic patterns

A lexico-syntactic pattern is a regular expression composed of words, grammatical or semantic categories, and symbols aiming to identify textual segments which match this expression. In the context of relation identification, the pattern characterizes a set of linguistic forms whose the interpretation is relatively stable and which corresponds to a semantic relation between terms (Rebeyrolle and Tanguy, 2000). Patterns are in fact very efficient, particularly in terms of precision, as they are adapted to the corpus. However, since their development is cost-expensive, it is conventional to implement generic patterns such as those of Hearst (Hearst, 1992). Here, we use a more complete list of 30 patterns from the work of Jacques and Aussenac (Jacques and Aussenac-Gilles, 2006)[5]. We have also extended this set of patterns with more specific (ad-hoc) patterns which better fit the template structure of disambiguation pages (Ghamnia, 2016). This set of enriched patterns are the one used in our experiments.

3.3.2 Distant supervision learning

We have chosen to use the principle of distant supervision proposed by Mintz and colleagues (Mintz et al., 2009). This approach consists in aligning an external knowledge base to a corpus and in using this alignment to learn relations. The learning ground is based on the hypothesis that "if two entities participate in a relation, all sentences that mention these two entities express that relation". Although this hypothesis seems too strong, Riedel and colleagues (Riedel et al., 2010) have showed that it makes sense when the knowledge base used to annotate the corpus is derived from the corpus itself.

As with any supervised learning method, it is necessary to create a set of examples, to train a statistical model on these examples, and to evaluate the model on a test set or by cross-validation. The originality of this approach refers to the fact that the learning examples are automatically built with the help of a semantic resource: the class associated to a pair of terms present in a same sentence, corresponds to the the relation (if it exists) that binds these terms in the external resource. Once trained from the learning examples, a multi-class classification algorithm makes it possible to associate a class (and therefore a relation) with each example of a new corpus.

We have adapted this method by focusing on the hypernym relation, and proceeding to a binary classification. A pair of terms is classified as a positive (negative) example if the two terms denoting two concepts that exist in the semantic resource are linked (are not linked) with the hypernymy relation

[4]http://babelnet.org/

[5]A JAPE implementation of these two types of patterns is available on https://github.com/aghamnia/SemPediaPatterns

in this resource. In all other cases, the term pair is not an example of learning. Our learning examples are constructed with reference to the semantic resource BabelNet which has the advantage of integrating various knowledge bases including DBPedia, the semantic resource that we want to enrich in a long term. In addition, the hypernymy relation is more straightforward expressed in BabelNet than in DBPedia.

Each example is built from a context which encompasses the two terms that are possibly linked by a relation. A context (or window) consists in n (n being the size of the window) tokens preceding, following and separating the two terms. The features are then extracted from that context. These features are described in Table 1.

Scope	Features	Signification	Type
Token	POS	Part Of Speech	string
	lemma	Lemmatized form of the token	string
Window	distT1	Number of tokens between the token and Term1	integer
	distT2	Number of tokens between the token and Term2	integer
	nbMotsFenłtre	Number of tokens in the window	integer
	distT1T2	Number of tokens between Term1 and Term2	integer
Sentence	nbMotsPhrase	Number of tokens in the sentence	integer
	presVerbe	Presence of a verbal form	boolean

Table 1: Features set.

Although the features we use here do not take into account more sophisticated structures, as dependency trees (as discussed in Section 2), they provide quite good results, as discussed in the next sections.

We illustrate the content of a feature vector with the following example where the length of the window is fixed to 3 (we have evaluated windows of dimensions 1, 3 and 5, the optimum being obtained for length 3):

"Lime ou citron vert, le fruit des limettiers : Citrus aurantiifolia et Citrus latifolia"

Mapping the list of terms leads to annotate the sentence with terms `Lime`, `citron`, `citron vert`, `vert`, `fruit`. Let us consider the pair <`Lime`, `fruit`> randomly chosen by the system: Term1=`Lime` and Term2=`fruit`.

The system thus extracts:

```
Terme1 ou citron vert, le Terme2 des limettiers :
```

where tokens corresponding to terms have been replaced with Term1 and Term2. TreeTagger annotation allows to replace the exact form of tokens by their part of speach followed by their lemma:

```
Terme1 KON/ou NOM/citron ADJ/vert PUN/, DET:ART/le Terme2
PRP:det/du NOM/limettier PUN/:
```

Finally, feature functions give distances (in number of tokens) between a token and the annotated terms in the form of the pair of values, the number of tokens between Term1 and Term2 (here 5) and the number of tokens in the whole sentence (here 16). The last feature indicates the presence of a verbal form to discriminate poorly-written text from well-written text.

```
(1,-5) (2,-4) (3,-3) (4,-2) (5,-1) (7,1) (8,2) (9,3) 5 16 true
```

The entire example leads to the following representation:

```
Terme1 KON/ou NOM/citron ADJ/vert PUN/, DET:ART/le Terme2
PRP:det/du NOM/limettier PUN/:
(1,-5) (2,-4) (3,-3) (4,-2) (5,-1) (7,1) (8,2) (9,3) 5 16 true
```

This example is a positive one as a hypernym link between *lime* and *fruit* exists in BabelNet.

From the whole set of examples produced according to the process described above, we randomly selected 3000 positive examples and 3000 negative examples (from a total of 84169 examples). From these 6000 examples, 4000 are used as the set of training examples and 2000 form for the test set (with a rate of 50% of positive examples, for both training and test sets). We are aware that the strategy we follow to split the set of examples may affect the results. Although alternative strategies consider, for instance, the zero-lexical overlapping, as adopted by Weeds and colleagues (Weeds et al., 2014) and Levy and colleagues (Levy et al., 2015), we can not follow this kind of strategy here due to the nature of the corpus, where each sentence of a page corresponds to a characterization or a definition of the entity described by this page.

We have trained a binary logistic regression algorithm, the Maximum Entropy classifier MaxEnt (Berger et al., 1996) on the training set. When applying this algorithm on the test set, we obtained a recall of 0.63 and an accuracy of 0.71.

4 Results and discussion

In the following, we discus the results of the approaches described above and we evaluate their complementarity, as well as the advantage of combining their results.

4.1 Results

The quantitative evaluation we present in this section is not intended to measure the performance of the approaches in absolute terms, but rather to know the order of magnitude of the number of relations found by each of them, whether they are common or specific. This evaluation is based on the reference corpus. The set of examples from the reference corpus contains 688 true positive examples (TP) and 267 true negative examples (TN). We consider the relations extracted by each of the approaches as well as the intersection and the union of the set of relations. Table 2 provides the results in terms of precision, recall, F-measure and accuracy. We can observe that we obtain the best values of F-measure when combining both the results of patterns and MaxEnt.

	Patterns	MaxEnt	Patterns inter MaxEnt	Patterns union MaxEnt
Precision	0.81	0.71	0.75	0.73
Recall	0.46	0.63	0.32	0.77
F-measure	0.53	0.67	0.45	0.75
Accuracy	0.54	0.55	0.43	0.63

Table 2: Evaluation of the approaches.

As we will better discuss in the next section, the two approaches do not often find the same relations, what corroborates their complementarity.

4.2 Discussion

We have carried out an analysis of the nature of the differences in the set of extracted relations from both approaches. We first counted the number of relations found by each approach individually, by both of them, or by none of the two (Table 3), with respect to the true positive relations extracted from the reference corpus.

Among the 221 TP relations found by the two approaches, few of them (9 relations) are expressed by the verb *to be*, as for instance, between the terms *macédoine* (*macedonia*) and *salade de fruits* (*fruit salad*) in the sentence "La macédoine est une salade de fruits ou de légumes" (*Macedonia is a salad of fruit or of vegetables*). Almost all other relations correspond to the pattern "X, Y" as in the sentence "Le cheval de Troie, un mythe grec" (*The Trojan horse, a Greek myth*).

	Number TP
Found by patterns AND MaxEnt	221
Found by patterns AND NOT by MaxEnt	96
Found by MaxEnt AND NOT by patterns	210
Found neither by MaxEnt nor by patterns	161

Table 3: Number of true positives (TP) found by the approaches.

From the 96 relations found by patterns and which were not identified by MaxEnt, 19 of them are expressed with the help of the verb *to be*, in particular when the relation is not expressed at the beginning of the sentence, as the relation between *Babel fish* and *espèce imaginaire* (*imaginary species*) in the sentence "Le poisson Babel ou Babel fish est une espèce imaginaire" (*Babel fish or Babel fish is an imaginary species*). Most of the remaining relations match again the pattern "X, Y". We observe as well that the cause of the silence of MaxEnt in this case may be some specific syntactic variations, such as the presence of dates between parentheses, different punctuation, etc. Indeed, our learning model is sensitive to these variations as sentences are very short and present strong regularities.

Among the 210 relations found by MaxEnt, and not found by patterns, we can observe that (i) many relations are expressed with the help of a lexical inclusion, as for instance in the noun phrase *gare de Paris Bastille* (*Paris Bastille railway station*) used to identify the relation *gare de Paris Bastille* (*Paris Bastille railway station*) is a *gare* (*railway station*); (ii) some relations are expressed with the help of a state verb, as for the relation between *aigle* (*eagle*) and *oiseaux* (*birds*) in the sentence "Aigle désigne en fran{cais certains grands oiseaux rapaces" (*Eagle designates some large birds*). We can notice as well that MaxEnt is able to identify the relations expressed in textual units containing a coordination, as the relation between *poisson Babel* (*Babel fish*) and *espèce imaginaire* (*imaginary specie*) in the sentence "Le poisson Babel ou Babel fish est une espèce imaginaire", or between *Louis Babel* and *explorateur* (*explorer*) in the sentence "Louis Babel, prêtre-missionnaire oblat et explorateur" (*Louis Babel, oblate missionary priest and explorer*). Finally, MaxEnt is also able to identify the relations within the text using formatting as in the relation between *arête* (*ridge*) and *barbe de l'épi* (*beard of the ear*) in the sentence "Arête, "barbe de l'épi"" (*Ridge, 'beard of the ear'*) or between *Aigle* and *chasseur de mines* (*mine hunter*) in the sentence "Aigle (M647), chasseur de mines" (*Eagle (M647), mine hunter*).

From the 161 true positive relations missed by both patterns and MaxEnt, 64 are expressed in sentences that contain parenthetical clauses which separate two terms, as in the sentence "Un Appelant (jansénisme) est, au XVIIIe siècle, un ecclésiastique qui appelle ... (*An Appellant (jansenism) is, during the XVIIIth century, an ecclesiastic who calls ...*) where the relation *Appelant* (*Appellant*) is a *ecclésiastique* (*ecclesiastic*) is not found. 55 of them correspond to the relations expressed by head modifier. We could not precisely identify the silence of MaxEnt in this case. The remaining 42 cases concern forms of expression not supported by the patterns and too scarce to be learned by MaxEnt, such as "X such as Y".

Furthermore, we could also observe that patterns were able to identify relations between common names, rather than between named entities, whereas MaxEnt mainly finds relations between named entities. The reason is that some patterns identify phrases that may not be annotated with the LCorpus terms.

In summary, these results corroborate the gain brought by the combination of complementary methods on the same corpus. Firstly, we could notice that MaxtEnt is able to identify hypernym relations within complex phrases or textual structures, such as vertical item lists, provided they appear with a minimal frequency. Secondly, the different occurrences of relations within the same sentence are identified by the two methods, as seen above through the example "Le poisson Babel ou Babel fish est une espèce imaginaire". In these experiments, patterns and MaxEnt are complementary in a proportion of ˜1/3 vs. 2/3.

4.3 DBPedia enrichment

In a last stage, we evaluated how much our approach could enrich DBPedia with the extracted relations. To do so, we manually checked their presence/absence in DBPedia. This verification had to be manual because the annotated terms come from LCorpus and may differ from the labels in DBPedia. We queried DBPedia to check if entities with labels close to *Term1* and *Term2* were linked by a path made of *rdf:type* and *rdf:subclassOf* relations. We set to 3 the maximum path length.

From the 688 TP in the reference corpus, 199 relations were not expressed in DBPedia. 103 of these 199 relations were identified by the learning approach and 42 of them were found by patterns. Considering the union of the results of the two approaches, 125 identified relations were not in DBPedia (20 relations belonging to the intersection of the individual results). Table 4 presents the rate of enrichment of DBPedia with respect to the relations identified by each approach and the union of their results. These figures confirm that the Wikipedia text, which is under-exploited by Wikipedia extractors, contains hypernym relations other than those found in structured elements (infobox, categories, etc.).

Method	Enrichment rate
Patterns	21%
MaxEnt	51%
Pattern union MaxEnt	63%

Table 4: DBPedia enrichment rate.

5 Conclusion and perspectives

The study reported in this paper led us to set up a methodology to compare two relation extraction approaches, in order to analyze their complementarity. The first results are encouraging and converge with the work of (Malaisé et al., 2004) (Granada, 2015) (Buitelaar et al., 2005). We plan to push this research further on in several directions. We want to integrate other methods, taking into account other textual elements, for example the system of (Kamel and Trojahn, 2016) that deals with vertical and regular enumerative structures, or the tools developed in (Granada, 2015). For improving the performance of each method, in addition to a better pattern encoding, we plan to add new features to the learning process. Moreover, the method will have to be tested on another corpus including other types of Wikipedia pages.

Ultimately, our ambition is to cross the methods so that the results of some serve as richer inputs to others, and thus improve their performance. The first step in this direction would be to annotate the corpus using patterns and tag it to signal whether a pattern is (or is not) recognized in the context of two terms, which would be a strong sign of the presence of the relation. This type of feature would allow the classifier to recognize several types of relations in addition to hypernymy.

Acknowledgement

The authors would like the Midi-Pyrénées (now Occitanie Pyrénées-Méditerranée) Region who funded the SemPedia project and Adel Ghamnia's Ph.D. grant.

References

Agichtein, E. and L. Gravano (2000). Snowball: Extracting relations from large plain-text collections. In *Proceedings of the 5th ACM conference on Digital libraries*, pp. 85–94. ACM.

Banko, M., M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni (2007). Open information extraction from the web. In *IJCAI*, Volume 7, pp. 2670–2676.

Ben Abacha, A. and P. Zweigenbaum (2011). A Hybrid Approach for the Extraction of Semantic Relations from MEDLINE Abstracts. In A. Gelbukh (Ed.), *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing, Tokyo, Japan, February 20-26*, pp. 139–150. Springer Berlin Heidelberg.

Berger, A. L., V. J. Della Pietra, and S. A. Della Pietra (1996). A maximum entropy approach to natural language processing. *Computational linguistics 22*(1), 39–71.

Berland, M. and E. Charniak (1999). Finding parts in very large corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 57–64. Association for Computational Linguistics.

Brin, S. (1998). Extracting patterns and relations from the world wide web. In *International Workshop on The World Wide Web and Databases*, pp. 172–183. Springer.

Buitelaar, P., P. Cimiano, and B. Magnini (2005). Ontology learning from text: An overview. In *Ontology Learning from Text: Methods, Evaluation and Applications*, pp. 3–12. IOS Press.

Bunescu, R. C. and R. J. Mooney (2005). A shortest path dependency kernel for relation extraction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pp. 724–731. Association for Computational Linguistics.

Etzioni, O., M. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates (2004). Web-scale information extraction in knowitall:(preliminary results). In *Proceedings of the 13th international conference on World Wide Web*, pp. 100–110. ACM.

Fabre, C., N. Hathout, L.-M. Ho-Dac, F. Morlane-Hondère, P. Muller, F. Sajous, L. Tanguy, and T. Van De Cruys (2014, June). Présentation de l'atelier SemDis 2014 : sémantique distributionnelle pour la substitution lexicale et l'exploration de corpus spécialisés. In *21e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2014)*, Marseille, France, pp. 196–205.

Fader, A., S. Soderland, and O. Etzioni (2011). Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1535–1545. Association for Computational Linguistics.

Ghamnia, A. (2016). Extraction de relations dhyperonymie partir de wikipdia. In *Actes de la confrence conjointe JEP-TALN-RECITAL 2016*.

Granada, R. L. (2015). *Evaluation of methods for taxonomic relation extraction from text*. Ph. D. thesis, Pontifícia Universidade Católica do Rio Grande do Sul.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, COLING '92, Stroudsburg, PA, USA, pp. 539–545. Association for Computational Linguistics.

Jacques, M.-P. and N. Aussenac-Gilles (2006). Variabilité des performances des outils de TAL et genre textuel. Cas des patrons lexico-syntaxiques. *Traitement Automatique des Langues, Non Thmatique 47*(1).

Kamel, M. and C. Trojahn (2016). Exploiter la structure discursive du texte pour valider les relations candidates d'hyperonymie issues de structures énumératives parallèles. In *IC 2016 : 27es Journées francophones d'Ingénierie des Connaissances, Montpellier, France, June 6-10, 2016.*, pp. 111–122.

Kazama, J. and K. Torisawa (2007). Exploiting wikipedia as external knowledge for named entity recognition. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 698–707.

Kotlerman, L., I. Dagan, I. Szpektor, and M. Zhitomirsky-geffet (2010, October). Directional distributional similarity for lexical inference. *Nat. Lang. Eng. 16*(4).

Lenci, A. and G. Benotto (2012). Identifying hypernyms in distributional semantic spaces. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pp. 75–79. Association for Computational Linguistics.

Levy, O., S. Remus, C. Biemann, and I. Dagan (2015). Do supervised distributional methods really learn lexical inference relations? In *HLT-NAACL*.

Malaisé, V., P. Zweigenbaum, and B. Bachimont (2004). Detecting semantic relations between terms in definitions. In *COLING 2004 CompuTerm 2004: 3rd International Workshop on Computational Terminology*, pp. 55–62. COLING.

Mintz, M., S. Bills, R. Snow, and D. Jurafsky (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pp. 1003–1011. Association for Computational Linguistics.

Morin, E. and C. Jacquemin (2004). Automatic acquisition and expansion of hypernym links. *Computers and the Humanities 38*(4), 363–396.

Morsey, M., J. Lehmann, S. Auer, C. Stadler, and S. Hellmann (2012). Dbpedia and the live extraction of structured data from wikipedia. *Program 46*(2), 157–181.

Nakashole, N., G. Weikum, and F. Suchanek (2012). Patty: A taxonomy of relational patterns with semantic types. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, Stroudsburg, PA, USA, pp. 1135–1145. Association for Computational Linguistics.

Navigli, R. and P. Velardi (2010). Learning word-class lattices for definition and hypernym extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, Stroudsburg, PA, USA, pp. 1318–1327. Association for Computational Linguistics.

Pantel, P. and M. Pennacchiotti (2008). Automatically harvesting and ontologizing semantic relations. *Ontology learning and population: Bridging the gap between text and knowledge*, 171–198.

Rebeyrolle, J. and L. Tanguy (2000). Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires. *Cahiers de Grammaire 25*, 153–174.

Riedel, S., L. Yao, and A. McCallum (2010). Modeling relations and their mentions without labeled text. In *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 148–163.

Riedel, S., L. Yao, A. Mccallum, and B. M Marlin (2013). Relation extraction with matrix factorization and universal schemas. In *Proceedings of NAACL-HLT 2013*.

Rodriguez-Ferreira, T., A. Rabadan, R. Hervas, and A. Diaz (2016, may). Improving Information Extraction from Wikipedia Texts using Basic English. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Santus, E., A. Lenci, Q. Lu, and S. Schulte im Walde (2014). Chasing hypernyms in vector spaces with entropy. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pp. 38–42. Association for Computational Linguistics.

Séguéla, P. and N. Aussenac-Gilles (1999). Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine. In *Conférence ingénierie des connaissances*, pp. 79–88.

Snow, R., D. Jurafsky, and A. Y. Ng (2004). Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems 17*.

Snow, R., D. Jurafsky, and A. Y. Ng (2006). Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pp. 801–808.

Suchanek, F. M., G. Kasneci, and G. Weikum (2007). Yago: A core of semantic knowledge unifying wordnet and wikipedia. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pp. 697–706.

Sumida, A. and K. Torisawa (2008). Hacking wikipedia for hyponymy relation acquisition. In *IJCNLP*, Volume 8, pp. 883–888.

Weeds, J., D. Clarke, J. Reffin, D. J. Weir, and B. Keller (2014). Learning to distinguish hypernyms and co-hyponyms. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pp. 2249–2259.

Yap, W. and T. Baldwin (2009). Experiments on pattern-based relation learning. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pp. 1657–1660. ACM.

Yates, A., M. Cafarella, M. Banko, O. Etzioni, M. Broadhead, and S. Soderland (2007). Textrunner: open information extraction on the web. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pp. 25–26. Association for Computational Linguistics.

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 112 (2017) 1764–1773

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017, 6-8 September 2017, Marseille, France

A Distant Learning Approach for Extracting Hypernym Relations from Wikipedia Disambiguation Pages

Mouna Kamel[a,b,*], Cassia Trojahn[b], Adel Ghamnia[b,c], Nathalie Aussenac-Gilles[b], Cécile Fabre[c]

[a] *Université de Perpigan, France*
[b] *IRIT, CNRS, Université de Toulouse, France*
[c] *CLLE, équipe ERSS, Université de Toulouse, France*

Abstract

Extracting hypernym relations from text is one of the key steps in the automated construction and enrichment of semantic resources. The state of the art offers a large varierty of methods (linguistic, statistical, learning based, hybrid). This variety could be an answer to the need to process each corpus or text fragment according to its specificities (e.g. domain granularity, nature, language, or target semantic resource). Moreover, hypernym relation may take different linguistic forms. The aim of this paper is to study the behaviour of a supervised learning approach to extract hypernym relations whatever the way they are expressed, and to evaluate its ability to capture regularities from the corpus, without human intervention. We apply a distant supervised learning algorithm on a sub-set of Wikipedia in French made of disambiguation pages where we manually annotated hypernym relations. The learned model obtained a F-measure of 0.67, outperforming lexico-syntactic pattern matching used as baseline.

© 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of KES International

Keywords: Distant learning, hypernym relation, knowledge extraction, Wikipedia, knowledge bases

1. Introduction

Extracting hypernym relations from text is one of the key steps in the automated construction and enrichment of semantic resources, since this kind of relation provides the hierarchical backbone structure which allows for entity type assignment. Several hypernym extraction methods have been proposed in the literature, trying to better identify the different ways this kind of relation is expressed in written natural language.

While linguistic methods rely on lexico-syntactic patterns used to identify clues of relation between terms[1], statistical methods, which are predominant, rely on distributional spaces[2] or on supervised[3] or non-supervised[4] learning methods. Most of these technique apply on well-written text (i.e., where the language syntactic structure is well-formed), without taking into account its structure. However, some of them exploit specific textual structures, e.g. definitions[5], enumerative structures[6] or elements like infoboxes, categories or links[7] in the case of Wikipedia.

* Corresponding author.
E-mail address: mouna.kamel@irit.fr

1877-0509 © 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of KES International
10.1016/j.procs.2017.08.208

These methods have been developed for corpora with their own specificities, e.g., domain granularity (general or specific), corpus gender (encyclopedic, scientific, journalistic, etc.), language, explicitness of the text structure (structured, semi-structured or unstructured). Another design parameter is the intended aim (i.e. extracting linguistic relations or formal triples, annotating text or populating a knowledge base). When relations have to be included in a semantic resource, the nature of this resource (thesauri, termino-ontologies or heavy ontologies) also impacts the extraction process. In addition, whatever the input corpus, hypernym relations may be expressed in different forms.

In this paper we study the behaviour of a distant supervised learning approach[8] on a corpus where the hypernym relation is expressed in different forms. Supervised learning algorithms carry out classification based on features of the entities to be classified. When applied to text, these features include various linguistic clues (either syntactic, semantic, lexical, visual, structural, distributional clues). Training the algorithm requires the corpus to be annotated with examples of class to be learned, which is complex and time consuming. Distant supervision overcomes this limitation by relying on an external semantic resource that is mapped to the corpus to automatically generate relation annotations. Being free of manual annotation, this approach can easily be applied to any corpus with regular structures.

This work is carried out within the SemPedia[1] project whose goal is to enrich DBPedia[2] for French, by specifying and implementing a set of new Wikipedia extractors dedicated to the hypernym relation. We focus on French because semantic resources targeting this language are scarce. The French DBPedia resource is 20,000 times poorer than DBPedia in English. Then, we built a corpus made of Wikipedia disambiguation pages in French. These pages list the Wikipedia articles whose title is ambiguous, and give a definition of each accepted meaning. Therefore, they are rich in named entities and in hypernym relations expressed through textual definitions and entity types. Moreover, we observed that pages are structured at various degrees depending on the language. For example, pages in English are very highly structured and mostly contain "low-written text" (e.g., opposed to well-written text), where the structure substitutes the lack of full syntax and expresses a good part of the text meaning. The French pages, on the contrary, more readily mix written text and low-written text. So they are a favorable case of relation-rich pages.

Concerning the automatic annotation, we used BabelNet[9], a knowledge resource which is (partially) derived from the training Wikipedia corpus, as recommended by the distant supervision approach. We compare our approach to a symbolic one based on lexico-syntactic patterns, which have been largely used to identify hypernym relations, in particular, on corpora rich in definitions[1], such as the corpus we used in our experiments.

This paper is structured as follows. Section 2 deals with the main related work. Section 3 introduces the distant supervised learning approach and the Maximum Entropy algorithm that we use. Then we present the hypernym relation learning task in Section 4. Section 5 describes the experiments and discusses the obtained results. Finally, Section 6 concludes the paper and draws perspectives for future work.

2. Related work

In the field of relation extraction, the pioneering work of the linguistic methods is that of Hearst[1] which defined a set of lexico-syntactic patterns specific to the hypernym relation for English. This work has been adapted to French in order to identify different types of relations[10], hypernym relations between terms[11] or meronymic relations[12]. Moreover paterns have been progressively learned from text thanks to learning techniques.

With respect to statistical approaches, Snow and colleagues[13] and Bunescu and colleagues[14] applied supervised learning techniques on a set of manually annotated examples. Because the cost of manual annotation is one of the main limitations of supervised learning, distant supervision learning consists in building the set of examples thanks to an external resource[8]. Distant supervision avoids the manual annotation phase by matching relations from the external knowledge base on the corpus. This process allows to automatically annotate relation occurrences that will become learning examples. Another way to avoid manual annotation has been proposed by Brin[15] who uses a selection of patterns to construct the set of examples, thanks to a semi-supervised learning method called bootstrapping. Agichtein and Gravano[16], and Etzioni and colleagues[17] have used this method and added semantic features to identify relations between named entities. IAn alternative technique is unsupervised learning based on clustering techniques. Yates and

[1] http://www.irit.fr/Sempedia

[2] DBPedia is a crowd-sourced community effort to extract structured information from Wikipedia and make it available on the Linked Open Data

colleagues[18] and Fader and colleagues[19] implemented unsupervised learning and used syntactic features to train their classifiers on relations between named entities. Some of these works are also based on distributional analyses[2,20]. To better understand how to take into account the specificities of the corpora, Yap and Baldwin[21] studied the impact of the corpus and the size of training sets on the performance of supervised methods for the extraction of different types of relation (hypernym, synonymy and antonymy). Granada[22] compared the performance of different methods (patterns-based, head-modifier, and distributional ones) for the task of hypernym relation extraction on various kinds of corpora (encyclopedic, journalistic) in several languages.

Relation extraction can also take advantage of the page layout in two different ways. The first one relies on documents written in a markup language. The semantics of the tags and their nested structure maybe exploited for the identification of relations. A collection of XML documents has been used to build ontologies[23,24], while a collection of HTML or MediaWiki documents has been processed to build taxonomies[25]. The second category of approaches bears on specific documents or parts of them, for which the layout defines a precise semantics, such as dictionaries and thesaurus[26] or specific and well localized textual structures such as tables, categories[27,28] or infoboxes[29] from Wikipedia pages. Any of these textual structures can also be made explicit thanks to a markup language.

Finally, as our study focuses on the Wikipedia corpus, the extracted relations could be exploited to enrich DBPedia. Several tools, called "extractors" have been developed to analyze each type of structured data in Wikipedia. Morsey and colleagues[7] developed 19 of such extractors that produce a formal representation of entities and relations identified within various structural elements from Wikipedia: abstracts, images, infobox, etc. Other works have targeted specific relations, mainly hypernym relations. For example, Suchanek and colleagues[28] used the 'Category' hierarchy of Wikipedia to build hypernym relations in the Yago knowledge base. Kazama and Torisawa[30] exploited the 'Definition' part of the pages, whereas Sumida and Torisawa[25] extracted knowledge from the menu items. Recent works proposed the automatic creation of MultiWiBi[31], an integrated bitaxonomy of Wikipedia pages and categories in multiple languages. Still, relation extraction from the text in Wikipedia pages has been little used to feed DBPedia[32]. Hence, most of the knowledge from these pages remains unexploited. This lack is even more important for pages in French.

This context led us to define methods that could extract knowledge from the text in French Wikipedia pages. As a first step towards this goal, we target the extraction of hypernym relations whatever the way they are expressed, be it in well-written or in low-written text. Given the size of the corpus and the cost and expertise by the manual annotation of examples, we apply a distant supervised learning algorithm. We also decided to combine various features, at different linguistic levels (from morphology and syntax to discourse and layout). This approach can be carried out on any corpus presenting structural and/or linguistic regularities, such as web documents, and it is language-independent.

3. Background

3.1. Distant supervision learning

Distant supervision learning[33,8] refers to learning algorithms where the training examples are automatically collected using a knowledge base. The set of examples is built by aligning the knowledge base to a corpus. The resulting alignments (or text annotations) and their features are then used to train the system and learn relations. The learning ground is based on the hypothesis that "if two entities participate in a relation, all sentences that mention these two entities express that relation". Although this hypothesis seems too strong, Riedel and colleagues[34] show that it makes sense when the knowledge base used to annotate the corpus is derived from the corpus itself. Mintz and colleagues[8] use Freebase as external resource. For every pair of entities linked in Freebase and appearing together within a sentence, a positive learning example is built, i.e., the learning features are extracted from the sentence and added to a feature vector for that entity pair. The set of feature vectors feed a multi-class logistic regression classifier. While Mintz and colleagues[8] consider several relations at once with a multi-class classifier, given the size of our corpus, we only focus on the hypernym relation and a binary logistic regression classifier, the Maximum Entropy Algorithm.

3.2. The Maximum Entropy algorithm

To perform the binary classification task (*isA* or *not-isA* classes), we chose the Maximum Entropy classifier (MaxEnt)[35] which is relevant when the conditional independence of the features can not be assured. This is particularly

true in NLP where features are usually words which obviously are not independent in their use (they are bound by syntactic and sematic rules). Furthermore, MaxEnt allows the management of a great number of features. It relies on the maximum entropy principle. Hence, it requires to define a set of constraints for each observation and to choose the distribution which maximizes the entropy while remaining consistent with the whole set of constraints[36]. In this context of optimisation under constraints, it is mathematically proved that a unique solution exists and that an iterative algorithm converges towards this solution[37].

The classical formula of MaxEnt is the following :

$$P(y|x) = \frac{1}{Z} \exp\left(\sum_i w_i f_i(x, y) \right)$$

where $P(y|x)$ gives the probability that the individual x (here a relation) belongs to the class y (here *isA* or *not-isA* classes). Each individual is encoded as a feature vector. The function f_i is a function called *feature* which determines the constraints of the model. The weights w_i associated to each feature account for the probability to belong to a class. Z is a normalization constant which ensures that the sum of probabilities of one individual is equal to 1.

To estimate the parameter values $\hat{w}$, we use the likelihood function that aims at determining the best estimators:

$$\hat{w} = argmax \sum_j log(P(y_j|x_j))$$

where the (x_j, y_j) belongs to the set of training data. In our work, we used the OpenNLP (version 1.5.0) implementation of the MaxEnt algorithm[3].

4. Hypernym relation learning task

In this section, we firstly describe the resources we used for training the models. Then, these models are presented.

4.1. Resources

We use the Wikipedia sub-corpus composed of all French disambiguation pages. These pages list the Wikipedia articles whose title is ambiguous, and give a definition of each accepted meaning. Therefore they are rich in named entities and in hypernym relations. Relations are expressed in different textual structures, usually established by the guidelines of the drafting charter. These HTML pages are semi-structured textual documents that combine different levels of text structuring, translated by typographical and layout features. The combination of these feature lead to a large variety of possibilities to express hypernym relations. Figure 1 presents the *Mercure* disambiguation page[4], where hypernym relations are expressed thanks to the following linguistic elements: the lexicon (*le mercure est un élément chimique*), the punctuation (the comma in *le Mercure, un fleuve du sud de l'Italie*), lexical inclusion (*appareil de mesure*, implying that *appareil de mesure* is an *appareil*), layout (disposition and typography) as used in enumerative structures (*la diode à vapeur de mercure est un appareil de mesure*, *la pile au mercure est un appareil de mesure*, etc.).

Depending on the language, the disambiguation pages are more or less structured. For instance, pages in English are very highly structured and contain essentially low-written text (noun-phrases and item lists). The French pages, in contrast, more readily mix the well-written text and the low-written text. So the French pages are particularly relevant for our experiment with a larger variety of ways to formulate hypernym relations.

The semantic resource that we use for building training examples is the BabelNet semantic network. BabelNet is both a multilingual encyclopedic dictionary, with lexicographic and encyclopedic coverage of terms, and a semantic network that connects concepts and named entities in a large network of semantic relations, made up of about 14 million entries. It has been automatically created by linking the encyclopedia Wikipedia to other resources, among which WordNet as source of lexical relations. In BabelNet, missing lexical entries in resource-poor languages have

[3] http://opennlp.apache.org/
[4] https://fr.wikipedia.org/wiki/Mercure

Physique et chimie [modifier | modifier le code]

- Le mercure (symbole Hg) est un élément chimique.
- Le terme mercure rouge désignait au XIXe siècle l'iodure de mercure. Dans la dernière partie du XXe siècle il a été appliqué à une substance imaginaire, présentée comme un matériau stratégique rentrant dans la construction des armes nucléaires.
- Le millimètre de mercure (symbole mmHg), ou torr, est unité de mesure de pression.
- Plusieurs appareils de mesure ou méthodes physiques font référence au mercure, dont notamment :
  - la diode à vapeur de mercure,
  - la pile au mercure,
  - la pompe à mercure,
  - le porosimètre à mercure (en),
  - le thermomètre à mercure.

Toponyme et hydronyme [modifier | modifier le code]

Mercure est un nom de lieu notamment porté par :

- Mercure, une station du métro de Lille Métropole ;
- le Mercure, un fleuve du sud de l'Italie ;
- les îles Mercure, un archipel néo-zélandais, au large de la péninsule de Coromandel.
- le lac Mercure, un lac de l'île principale de l'archipel des Kerguelen, dans les Terres australes et antarctiques françaises ;
- le monastère Saint-Mercure, un important monastère féminin copte orthodoxe, situé dans le vieux Caire (Égypte) ;
- le mont Mercure, une montagne d'Italie ;
- Saint-Michel-Mont-Mercure, une ancienne commune française située dans le département de la Vendée, en région Pays-de-la-Loire
- la Vallée du Mercure, un grand bassin fluvial italien situé dans le sud de la Basilicate et le nord de la Calabre, et qui fut occupé par u lac au Pliocène.

Fig. 1. Excerpt from the French disambiguation page of the word *Mercure*

been filled with the help of statistical machine translation. As a consequence, the hypernym relation is well covered. Moreover, as long as BabelNet is derived from Wikipedia, the training knowledge base is derived from the training text, as recommended by the distant supervision approach.

We divided the set of disambiguation pages into two sets: a reference corpus composed of 20 pages, and a training corpus composed of the remaining pages (5904 pages). The two corpora have been pre-processed: (i) the plain text has been extracted from these pages, with the help of a WikiExtractor[5]; (ii) this plain text has been annotated with Part-Of-Speech and lemma (using TreeTagger); and (iii) it has been annotated with terms (including both single words and multiword expressions) corresponding to labels of concepts in BabelNet for the training corpus (resp. with the set of manually annotated relations for the reference corpus).

4.2. Features, examples and training models

4.2.1. Features

For each sentence in the training set, we extract a set of features from a window of size n. A sentence in the training text contains two terms tagged $Term1$ and $Term2$, that result from the annotation with the BabelNet concept labels. The window includes n tokens before $Term1$, $Term1$, the tokens between $Term1$ and $Term2$, $Term2$, n tokens after $Term2$. Only the features of the tokens belonging to this window are processed. We consider three kinds of features: those involving tokens, those involving sentences and those involving sentence windows. Currently, we focus on lexical and grammatical features, and some heuristics inspired by the work of Lin and colleagues[38], as they seem enough to provide good results. In fact, we decided no to use syntactic features because syntactic parsers provide poor results on low-written text. Furthermore, limiting the number of required NLP tools makes the approach easier to reproduce, especially for languages for which such tools are scarce. Table 1 presents the set of selected features.

4.2.2. Learning examples

Given that we only consider sentences that contain at least two terms denoting two concepts in BabelNet, the construction of training examples leans on the following assumptions:

[5] http://wiki.apertium.org/wiki/Wikipedia_Extractor

Scope	Features	Signification	Type
Token	POS	Part Of Speech	string
	lemma	Lemmatized form of the token	string
	distT1	Number of tokens between the token and Term1	integer
	distT2	Number of tokens between the token and Term2	integer
Window	nbWordsWindow	Number of tokens in the window	integer
	distT1T2	Number of tokens between Term1 and Term2	integer
Sentence	nbWordsSentence	Number of tokens in the sentence	integer
	presVerb	Presence of a verbal form	boolean

Table 1. Features set.

1. If two terms (in the same sentence) denote two concepts linked by a hypernym relation in BabelNet, a positive example will be built from the sentence window encompassing these two terms;
2. If no hypernym relation links the two BabelNet concepts whose labels occur in the same sentence, a negative example will be built from the sentence window encompassing these two terms.

Hence, for each sentence of each page of the training corpus, we randomly choose a pair of labeled terms[6] $Term1$ and $Term2$ occurring in that sentence. We then explore the BabelNet hierarchy to check whether $Term1$ and $Term2$ denote concepts linked by the hypernym relation. We empirically set a maximum length path of 3 levels in the hierarchy[39]. We give below an example of a feature vector, with a window size set to 3:

"Lime ou citron vert, le fruit des limettiers : Citrus aurantiifolia et Citrus latifolia"

Matching the BabelNet list of terms lead to annotate the sentence with the terms `Lime`, `citron`, `citron vert`, `vert`, `fruit`. Let us consider the pair <`Lime`, `fruit`> randomly chosen by the system : Term1=`Lime` and Term2=`fruit`. The system thus extracts:

```
Terme1 ou citron vert, le Terme2 des limettiers :
```

where tokens corresponding to terms have been replaced with $Term1$ and $Term2$. Tree Tagger parsing allows to replace the exact form of tokens by their part-of-speech followed by their lemma:

```
Terme1 KON/ou NOM/citron ADJ/vert PUN/, DET:ART/le Terme2 PRP:det/du NOM/limettier PUN/:
```

We finally compute distance features: for each token in the window, the feature is a pair of values representing the distance (in number of words) of this token respectively with $Term1$ and $Term2$. The last three features are the number of tokens between $Term1$ and $Term2$ (here 5); the number of tokens in the whole sentence (here 16); true or false depending on the presence (or absence) of a verbal form. This feature contributes to discriminate low-written text from well-written text.

```
(0,6) (-1,5) (-2,4) (-3,3) (-4,2) (-5,1) (-6,0) (-7,-1) (-8,-2) (-9,-3) 5 16 false
```

Here is the entire feature list for this example :

```
Terme1 KON/ou NOM/citron ADJ/vert PUN/, DET:ART/le Terme2 PRP:det/du NOM/limettier PUN/:
(0,6) (-1,5) (-2,4) (-3,3) (-4,2) (-5,1) (-6,0) (-7,-1) (-8,-2) (-9,-3) 5 16 false
```

This example is a positive one as a hypernym link between "lime" and "fruit" exists in BabelNet.

4.2.3. Training models

From the whole set of 84169 training examples produced according to the process described above, 4792 examples are labeled as positive, and 79377 are labeled as negative. We randomly took 3000 positive examples and 3000

[6] In order to obtain reasonable computation time and to provide sets of examples of reasonable size, we do not compute all possible combinations of pairs of terms from a single sentence

negative examples. From these 6000 examples, 4000 were used as the set of training examples and 2000 as the test set (with a rate of 50% of positive examples, for both training and test sets).

We then produced different learning models (according to the supported features and with different sizes of sentence windows). We evaluated them in terms of precision, recall and F-measure, as shown in Table 2. We can observe that the best results were obtained for a window of size 3. This result can be explained on the one hand by the fact that with a window of size 1, we loose contextual information. On the other hand, a window of size 5 does not bring so much given that the corpus sentences are relatively short and low-written. Hence the contextual information obtained for that 5 word window is not discriminating.

Features : POS and Lemma			
	window size=1	window size=3	window size=5
Precision	0.67	0.69	0.69
Recall	0.72	0.78	0.76
F-measure	0.69	0.73	0.72

Table 2. Results for POS and Lemma features, for different window sizes.

From these results, we have trained another model using 3 as window size and all the features listed in Table 1. The results are reported in Table 3.

	Features : all features listed in Table 1 and window size=3
Precision	0.72
Recall	0.66
F-measure	0.69

Table 3. Results for all features of Table 1 and for a window size = 3.

Unlike what we expected, the model using only the POS and lemma with window of 3 (model called in the following `model_POSL`) outperforms also the model considering all the features listed in Table 1 and which we refer to as `model_AllFeatures` in the following, in terms of F-measure. This can be explained by the fact that for that kind of corpus, POS and lemma features seem to be discriminant. Nevertheless we decided to evaluate both models on the reference corpus.

5. Evaluation

The evaluation we present in this section aims at showing the performance (in terms of precision and recall) of the two learning models presented in the previous section (`model_POSL` and `model_AllFeatures`), when applied to a corpus with the same characteristics as the ones on which the models were trained. We also compare the results to two baselines, both based on lexico-syntactic patterns largely used in the literature to identify hypernym relations.

This evaluation concerns the reference corpus composed of 20 French disambiguation pages. Two annotators have annotated this corpus in a double blind process: 688 sentences contained 2 BabelNet terms have been annotated as true positives (hypernym relations) and 278 such sentences as true negatives (absence of hypernym relation). After the calculation of the annotation agreement (of 0.8) between annotators, all conflicts have been identified and resolved.

This corpus has been pre-processed as described in section 4.1. From sentences containing the manually annotated relations, we extracted feature vectors and submitted them to the two classifiers `model_POSL` and `model_AllFeatures`.

5.1. Baseline

As stated above, two baselines based on lexico-syntactic patterns have been used. A lexico-syntactic pattern is a regular expression composed of words, POS or semantic categories, and symbols aiming at identifying textual segments which match this expression. In the context of relation extraction, the pattern characterizes a set of linguistic forms whose interpretation is relatively stable and corresponds to a semantic relation between terms[40]. Patterns are

in fact very efficient, in particular in terms of precision, when they are adapted to the corpus. However, since their development is expensive, it is conventional to implement generic patterns such as those of Hearst[1]. We chose a more complete list of 30 patterns from the work of Jacques and Aussenac[41]. This set of patterns is our first baseline (called `Baseline 1` in the following).

In a second step, in order to better take into account the specificity of the corpus, we have defined ad-hoc patterns adapted to the low-written parts of the disambiguation pages[42]. The set of generic patterns together with the specific patterns is our second baseline[7] (called `Baseline 2` in the following).

5.2. Results and discussion

Table 4 shows the results of the two baselines and the two classifiers, in terms of precision, recall, F-measure and accuracy. We can observe that the `model_AllFeatures` model gives the best results, with a F-measure equal to that of `model_POSL` but with a better accuracy. As expected, in terms of precision, the generic patterns (`Baseline 1`) outperform all the other models, in detriment of recall. When using specific patterns (`Baseline 2`), recall is highly improved. However, they introduce some noise (false positives) and decrease precision.

	`Baseline 1`	`Baseline 2`	`Model_POSL`	`Model_AllFeatures`
Precision	0.96	0.81	0.68	0.71
Recall	0.04	0.46	0.66	0.63
F-measure	0.07	0.59	0.67	0.67
Accuracy	0.31	0.54	0.53	0.55

Table 4. Results for the two baselines and the two classifiers on the reference corpus.

Since the accuracy is substantially the same (except to `Baseline 1`), another way of looking at these results is given in Table 5, which presents the number of true positive hypernym relations per type of hypernym expression in the text, found by the baselines and by the classifiers. With the classifiers, we are able to identity the most recurrent forms of expression of the hypernym relation, namely those relations expressed with well-written text, as in "$Term1$ *is a* $Term2$", those expressed using layout and therefore expressed with low-written text, as in "$Term1$, $Term2$", and as well as those which can be identified thanks to the head-modifier method. While the learning approach is able to identify a larger variety of expression forms of the relation, outperforming the symbolic approach, the patterns introduced in the `Baseline 2` seem to fit well the corpus regularities.

	Relations expressed in well-written text	Relations expressed in low-written text	Relations expressed with head-modifiers	Total
`Baseline 1`	24	2	0	26
`Baseline 2`	23	294	0	317
`Model_POSL`	22	243	187	452
`Model_AllFeature`	31	261	139	431

Table 5. Results for the two baselines and the two classifiers for the different ways of expressing the hypernym relation.

The intersection of the results provided by these methods shows that 11 true positive hypernym relations were found both by `Baseline 1` and by the two classifiers, while 221 true positive hypernym relations were found both by `Baseline 2` and by the two classifiers. From a quantitative point of view, automatic learning identifies more examples than patterns, without any development cost, ensuring a systematic and less empirical approach. From a qualitative point of view, classifiers do not perform as well as patterns when relations are regularly expressed in the same way, as somehow expected, but they can identify more varied forms of relation expressions. For instance, patterns have not the ability to identify relations with head modifiers, while the classifiers are able to do so. The following examples show the expression of hypernym relations in complex sentences that could be correctly identified by the classifiers:

[7] A JAPE implementation of these two types of patterns is visible on the site:https://github.com/aghamnia/SemPediaPatterns

(1) <Louis Label, prêtre-missionnaire oblat> and <Louis Label, explorateur du Nouveau-Québec> in the sentence *Louis Babel, prêtre-missionnaire oblat et explorateur du Nouveau-Québec (1826-1912).*

(2) <fontaine, robinet de cuivre> in the sentence *La fontaine a aussi désigné le "vaisseau de cuivre ou de quelque autre métal, où l'on garde de l'eau dans les maisons", et encore le robinet de cuivre par où coule l'eau d'une fontaine, ou le vin d'un tonneau, ou quelque autre liqueur que ce soit.*

In these examples, the relations are expressed within textual units using conjunction, as in example (1), or with $Term1$ and $Term2$ being relatively far from each other in the sentence, as in example (2). A pattern approach would require the definition of new patterns to fit these cases.

6. Conclusion and perspectives

This paper has presented a distant supervision learning approach for extracting hypernym relations from a corpus where hypernymy is expressed in a variety of forms. We composed a corpus from Wikipedia disambiguation pages, which are rich in hypernym relations expressed through textual definitions and named entities. We have mainly used lexical and grammatical features. Even with this reduced set, we could observe that the learning approach is able to extract regularities from the corpus. It was able to correctly identify different ways of expressing relations, including a set of those that could be identified by patterns or head-modifiers, for instance.

We have evaluated learning and patterns independently. However their combination seems to be a good strategy. In fact, combining learning with a broad set of patterns, best suited to the corpus, could ensure the best results (at the cost of developing such patterns). In the case of very regular pages with singularities (such as disambiguation pages), the development of ad-hoc patterns is immediate and justified.

As future work, we plan to train a model on the whole set of Wikipedia pages. We also intend to investigate additional features such as semantic, distributional or lay-out features. Moreover, we plan to study how patterns and learning approaches can be better combined.

Acknowledgement

The Sempedia Project is funded by the French region Occitanie-Mediterranée-Pyrénées from 2015 to 2018. This work also contributed to the SparkinData project funded by the French FUI program from 2015 to 2017.

References

1. Hearst, M.A.. Automatic acquisition of hyponyms from large text corpora. In: *Proceedings of the 14th conference on Computational linguistics*. Association for Computational Linguistics; 1992, p. 539–545.
2. Lenci, A., Benotto, G.. Identifying hypernyms in distributional semantic spaces. In: *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*. Association for Computational Linguistics; 2012, p. 75–79.
3. Pantel, P., Pennacchiotti, M.. Automatically harvesting and ontologizing semantic relations. *Ontology learning and population: Bridging the gap between text and knowledge* 2008;:171–198.
4. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.. Open information extraction from the web. In: *IJCAI*; vol. 7. 2007, p. 2670–2676.
5. Malaisé, V., Zweigenbaum, P., Bachimont, B.. Detecting semantic relations between terms in definitions. In: Ananadiou, S., Zweigenbaum, P., editors. *COLING 2004 CompuTerm 2004: 3rd International Workshop on Computational Terminology*. Geneva, Switzerland: COLING; 2004, p. 55–62.
6. Fauconnier, J.P., Kamel, M.. Discovering Hypernymy Relations using Text Layout (regular paper). In: *Joint Conference on Lexical and Computational Semantics (SEM), Denver, Colorado, 04/06/2015-05/06/2015*. Association for Computational Linguistics (ACL); 2015, p. 249–258.
7. Morsey, M., Lehmann, J., Auer, S., Stadler, C., Hellmann, S.. Dbpedia and the live extraction of structured data from wikipedia. *Program* 2012;**46**(2):157–181.
8. Mintz, M., Bills, S., Snow, R., Jurafsky, D.. Distant supervision for relation extraction without labeled data. In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. 2009, p. 1003–1011.
9. Navigli, R., Ponzetto, S.P.. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* 2012;**193**:217–250.

10. Séguéla, P., Aussenac-Gilles, N.. Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine. In: *Conférence ingénierie des connaissances*. 1999, p. 79–88.
11. Morin, E., Jacquemin, C.. Automatic acquisition and expansion of hypernym links. *Computers and the Humanities* 2004;**38**(4):363–396.
12. Berland, M., Charniak, E.. Finding parts in very large corpora. In: *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. Association for Computational Linguistics; 1999, p. 57–64.
13. Snow, R., Jurafsky, D., Ng, A.Y.. Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems 17* 2004;.
14. Bunescu, R.C., Mooney, R.J.. A shortest path dependency kernel for relation extraction. In: *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics; 2005, p. 724–731.
15. Brin, S.. Extracting patterns and relations from the world wide web. In: *International Workshop on The World Wide Web and Databases*. Springer; 1998, p. 172–183.
16. Agichtein, E., Gravano, L.. Snowball: Extracting relations from large plain-text collections. In: *Proceedings of the fifth ACM conference on Digital libraries*. ACM; 2000, p. 85–94.
17. Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.M., Shaked, T., et al. Web-scale information extraction in knowitall:(preliminary results). In: *Proceedings of the 13th international conference on World Wide Web*. ACM; 2004, p. 100–110.
18. Yates, A., Cafarella, M., Banko, M., Etzioni, O., Broadhead, M., Soderland, S.. Textrunner: open information extraction on the web. In: *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. Association for Computational Linguistics; 2007, p. 25–26.
19. Fader, A., Soderland, S., Etzioni, O.. Identifying relations for open information extraction. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics; 2011, p. 1535–1545.
20. Fabre, C., Hathout, N., Ho-Dac, L.M., Morlane-Hondère, F., Muller, P., Sajous, F., et al. Présentation de l'atelier SemDis 2014 : sémantique distributionnelle pour la substitution lexicale et l'exploration de corpus spécialisés. In: *21e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2014)*. Marseille, France; 2014, p. 196–205.
21. Yap, W., Baldwin, T.. Experiments on pattern-based relation learning. In: *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. ACM; 2009, p. 1657–1660.
22. Granada, R.L.. *Evaluation of methods for taxonomic relation extraction from text*. Ph.D. thesis; Pontifícia Universidade Católica do Rio Grande do Sul; 2015.
23. Kamel, M., Aussenac-Gilles, N.. How can document structure improve ontology learning? In: *Workshop on Semantic Annotation and Knowledge Markup collocated with K-CAP*. 2009, .
24. O'Connor, M.J., Das, A.. Acquiring owl ontologies from xml documents. In: *Proceedings of the Sixth International Conference on Knowledge Capture*. New York, NY, USA: ACM. ISBN 978-1-4503-0396-5; 2011, p. 17–24.
25. Sumida, A., Torisawa, K.. Hacking wikipedia for hyponymy relation acquisition. In: *IJCNLP*; vol. 8. Citeseer; 2008, p. 883–888.
26. Jannink, J.. Thesaurus entry extraction from an on-line dictionary. In: *Proceedings of Fusion*; vol. 99. Citeseer; 1999, .
27. Chernov, S., Iofciu, T., Nejdl, W., Zhou, X.. Extracting semantics relationships between wikipedia categories. *SemWiki* 2006;**206**.
28. Suchanek, F.M., Kasneci, G., Weikum, G.. Yago: a core of semantic knowledge. In: *Proceedings of the 16th international conference on World Wide Web*. ACM; 2007, p. 697–706.
29. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.. Dbpedia: A nucleus for a web of open data. In: *The semantic web*. Springer; 2007, p. 722–735.
30. Kazama, J., Torisawa, K.. Exploiting wikipedia as external knowledge for named entity recognition. In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 2007, p. 698–707.
31. Flati, T., Vannella, D., Pasini, T., Navigli, R.. Multiwibi: The multilingual wikipedia bitaxonomy project. *Artificial Intelligence* 2016; **241**(Complete):66–102. doi:10.1016/j.artint.2016.08.004.
32. Rodriguez-Ferreira, T., Rabadan, A., Hervas, R., Diaz, A.. Improving information extraction from wikipedia texts using basic english. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Paris, France: European Language Resources Association (ELRA); 2016, .
33. Bunescu, R.C., Mooney, R.J.. Learning to extract relations from the web using minimal supervision. In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*. Prague, Czech Republic; 2007, .
34. Riedel, S., Yao, L., McCallum, A.. Modeling relations and their mentions without labeled text. In: *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases: Part III*; ECML PKDD'10. Berlin, Heidelberg: Springer-Verlag; 2010, p. 148–163.
35. Berger, A.L., Pietra, V.J.D., Pietra, S.A.D.. A maximum entropy approach to natural language processing. *Computational linguistics* 1996; **22**(1):39–71.
36. Jaynes, E.. Information theory and statistical mechanics. *Physical review* 1957;**106**(4):620.
37. Ratnaparkhi, A.. *Maximum entropy models for natural language ambiguity resolution*. Ph.D. thesis; University of Pennsylvania; 1998.
38. Lin, Y., Shen, S., Liu, Z., Luan, H., Sun, M.. Neural relation extraction with selective attention over instances. In: *ACL*. 2016, .
39. Kamel, M., Trojahn, C.. Exploiter la structure discursive du texte pour valider les relations candidates d'hyperonymie issues de structures énumératives parallèles. In: *IC 2016 : 27es Journées francophones d'Ingénierie des Connaissances (Proceedings of the 27th French Knowledge Engineering Conference), Montpellier, France, June 6-10, 2016*. 2016, p. 111–122.
40. Rebeyrolle, J., Tanguy, L.. Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires. *Cahiers de Grammaire* 2000;**25**:153–174. URL: `https://halshs.archives-ouvertes.fr/halshs-01322256`.
41. Jacques, M.P., Aussenac-Gilles, N.. Variabilit des performances des outils de TAL et genre textuel. Cas des patrons lexico-syntaxiques. *Traitement Automatique des Langues, Non Thmatique* 2006;**47**(1):(en ligne).
42. Ghamnia, A.. Extraction de relations dhyperonymie partir de wikipdia. In: *Actes de la confrence conjointe JEP-TALN-RECITAL*. 2016, .

6 Enrichissement de traits pour l'extraction de relations d'hyperonymie

Sommaire

6.1 Extraire des relations à partir de structures énumératives : motivations

La majorité des méthodes d'extraction de relations sémantiques sont appliquées sur du texte rédigé pour extraire des relations entre concepts et termes, sans tenir en compte la richesse des structures textuelles en relations comme les définitions [Malaisé et al., 2004], les structures énumératives [Fauconnier and Kamel, 2015] ou les liens hypertextes [Morsey et al., 2012] dans le cas de Wikipedia. L'originalité de notre travail dans les chapitres précédents est de tester l'hypothèse (déjà présente avec la notion *d'extracteur* dans Yago et DbPedia) qu'il faut multiplier les techniques et algorithmes pour analyser et traiter chaque type d'expression des relations en corpus : texte rédigé, définitions, mais aussi texte mis en forme (énumérations, listes) ou encore formats demi-structurés : tableaux, liens de référence, hiérarchies etc.

Cette dernière étude se focalise sur les structures énumératives verticales, parties de textes particulièrement riches en relations hiérarchiques hyperonymiques. Ces parties de textes sont d'excellentes sources de relations et on en trouve un grand nombre dans Wikipedia. Ces structures peuvent être formalisées par des relations de discours tirées de la théorie sur la structure du discours telle que la DRT de Asher [1993]. Ce travail fait suite à la thèse de Fauconnier [2016].

6.2 Enrichir une base de connaissances : difficultés soulevées

Cette dernière étude de la thèse se focalise sur l'exploitation des triplets tirés des textes pour enrichir la base de connaissances DBPedia en français, qui était notre objectif au moment de définir SemPedia, ou la ressource BabelNet, puisque nous l'avons utilisée pour annoter automatiquement les textes. Les relations d'hyperonymie entre termes peuvent donner lieu à des relations de spécialisation entre classes (représentées à l'aide de la propriété owl :SubClassOf dans les bases de connaissances) ou à des relations entre une entité spécifique et la classe à laquelle on peut la rattacher (représentées par la propriété rdfs :type). Donc une des sous-tâches du processus est de décider comment représenter la relation elle-même.

En amont de cet ajout à la base de connaissances, pour chaque triplet extrait de textes mentionnant des termes (souvent exprimés par des syntagmes nominaux), on doit décider de la représentation de ces termes : ils peuvent renvoyer à une des classes ou une des entités existantes, dont le terme est déjà une des étiquettes ou dont il va devenir une nouvelle étiquette ; mais le terme peut donner lieu à la création de nouvelles entités ou classes, qu'il faut situer "au mieux" dans la base de connaissances. Nous retrouvons ces deux tâches dans l'approche présentée ici.

Cette approche se veut indépendante de la langue et ne nécessite pas l'annotation manuelle d'exemples pour entraîner le système.

6.3 Synthèse de l'approche et des résultats

Pour enrichir une base de connaissances à l'aide de relations d'hyperonymie tirées de structures énumératives verticales, nous avons défini un processus en quatre étapes, décrit dans la première partie de l'article :

1. Modification de la structure énumérative pour produire des "phrases" comprenant les deux arguments de la relation. Pour cela, l'item mentionnés dans l'amorce de la structure énumérative est distribuée sur chacun des items de l'énumération. Seule la première proposition (unité textuelle) de l'item énumérative est considérée.
2. annotation des phrase à l'aide de termes du domaine tirés d'une ressource sémantique ;
3. sélection d'un couple de termes par item réécrit. Le premier terme du couple est celui présent dans l'amorce qui a été ajoutée au début de chaque item. Le choix du deuxième argument répond à l'application d'heuristiques pour retenir les termes le plus plausible au début de la suite de l'item. Parmi tous les couples possibles présents dans l'item, ceux retenus pour définir autant d'exemples positifs sont les couples pour lesquels la relation existe dans BabelNet.
4. représentation de chaque exemple sous forme d'un vecteur de traits selon le principe décrit au chapitre 4. Cependant, la liste des traits a été revue et enrichie pour obtenu un ensemble plus précis, cf la table 1 dans la partie 3.1 de l'article.

Les données pratiquement utilisées dans l'expérimentation évaluant l'approche sont décrites dans la partie 4 de l'article :

Ressource enrichie : Ce travail se focalise sur l'enrichissement de DBpédia.

Corpus : il s'agit d'un corpus extrait de Wikipedia formé de soit 131 500 exemples tirés de près de 4000 structures énumératives.

Construction des exemples : L'approche se veut indépendante de toute annotation manuelle : les exemples utilisés pour entraîner le modèle d'apprentissage sont produits par supervision distante à l'aide de la ressource BabelNet. Les termes de BabelNet sont projetés sur le corpus des énumérations et permettent de focaliser la recherche de relations sur les phrases contenant deux termes ainsi reconnus.

Méthode d'extraction : Le modèle d'apprentissage entraîné est également celui choisi au chapitre 4 : c'est un algorithme de régression Maximum Entropie (MaxEnt) appliqué aux vecteurs de traits représentant les exemples. AInsi, il est adapté au corpus. Cet algorithme classe dans la classe "relation d'hyperonymie" ou non tout nouveau vecteur de traits construit à partir d'une phrase contenant un couple de termes.

Entraînement du modèle d'apprentissage : Le corpus des exemples comporte une énorme majorité d'exemples négatifs (131404 contre 2766 positifs). L'ensemble retenu pour entraîner l'algorithme ne contient que 1844 exemples positifs et autant de négatifs.

Evaluation : Nous avons utilisé un corpus de référence construit dans notre équipe pour lequel les relations présentes dans les énumérations ont été annotées.

Résultats : Notre approche obtient une F-mesure de 0,71 ou 0,73 suivant la partie du corpus testée, et de situe au-dessus de l'algorithme servant de base (score de 0,6). Parmi celles-ci, 307 relations, près de 98% comportent deux termes qui ne son t pas présents dans DBPedia, et la relation d'hyperonymie elle-même n'est pas présente. Seules 2 relations sont présentes dans DBpedia avec leurs deux arguments.

Notre travail apporte donc bien une réponse à notre objectif initial, qui était d'enrichir DBpedia en français à l'aide de nouvelles connaissances tirées du texte des pages Wikipedia. Les resultats sont très encourageants malgré la qualité non encore optimale des algorithmes d'extraction de relations mis en oeuvre, et malgré une exploitation encore partielle de ces textes. Ils nous encouragent à traiter d'autres types de parties des pages riches en relations de définition.

6.4 Article tiré de cette étude

Kamel, A., Trojahn, C. (2018). Towards Enriching DBpedia from Vertical Enumerative Structures Using a Distant Learning Approach. In *Proceedings of the International Conference on Knowledge Engineering and Knowledge Management (EKAW 2018)*, Nov 2018, Nancy, France. pp.179-194.

Nous reproduisons cet article dans les pages qui suivent.

Towards Enriching DBpedia from Vertical Enumerative Structures Using a Distant Learning Approach

Mouna Kamel and Cassia Trojahn( )

Institut de Recherche en Informatique de Toulouse, Toulouse, France
{prenom.nom,cassia.trojahn}@irit.fr

Abstract. Automatic construction of semantic resources at large scale usually relies on general purpose corpora as Wikipedia. This resource, by nature rich in encyclopedic knowledge, exposes part of this knowledge with strongly structured elements (infoboxes, categories, etc.). Several extractors have targeted these structures in order to enrich or to populate semantic resources as DBpedia, YAGO or BabelNet. The remain semi-structured textual structures, such as vertical enumerative structures (those using typographic and dispositional layout) have been however under-exploited. However, frequent in corpora, they are rich sources of specific semantic relations, such as hypernyms. This paper presents a distant learning approach for extracting hypernym relations from vertical enumerative structures of Wikipedia, with the aim of enriching DBpedia. Our relation extraction approach achieves an overall precision of 62%, and 99% of the extracted relations can enrich DBpedia, with respect to a reference corpus.

1 Introduction

In many fields such as artificial intelligence, semantic web or question answering, applications require a reasoning ability, based on semantic resources that describe concepts and relations between. Manually constructing this kind of resource is cost-intensive and results in domain-specific resources of low coverage. However, more than ever automated support for large scale construction of such resources becomes essential. This involves automatically extracting relations from text for building, enriching or populating them. This task usually relies on general purpose corpora as Wikipedia or WordNet and on knowledge extractors mainly exploiting their specific structural elements [22] or sub-corpora [15]. Several of these extractors have targeted these structures in order to enrich or to populate resources as DBpedia [2], YAGO or BabelNet [23].

Enriching DBpedia means identifying new semantic relations from Wikipedia pages. A Wikipedia page is composed of different textual structures which can be divided into three main categories: strongly structured elements, paragraphs which contain plain text, and semi-structured textual units. Strongly structured elements such as infoboxes or User Generated Categories (UGCs) benefit from

https://doi.org/10.1007/978-3-030-03667-6_12

a strong layout, convey a well defined semantics and contain poor written text. Extractors exploiting these elements usually focus on relations (*birthPlace*, *birthDate*, *win-prize*, etc.) which are mostly limited to named entities such as cities, persons, species, etc. With respect to plain text, it has been exploited by numerous relation extraction systems, more often abstracts (whose first sentence is a definition) for identifying hypernym relations[1], other paragraphs for identifying relations in a context of Open Information Extraction. Wikipedia pages are also composed of textual structures, such as titles, subtitles, vertical enumerative structures (i.e. enumerations using typographical and dispositional markers (Fig. 1)). We consider these textual structures as semi-structured ones, as they have the particularity to combine well-written text and layout. Although they express relations which are more often hierarchical relations, these types of structures remain under-exploited as they can not be correctly processed by most classical NLP tools.

The aim of this paper is to show to what extent DBpedia may be enriched with hypernym relations extracted from vertical enumerative structures (VES) present in Wikipedia pages. This kind of relation is central to the construction and enrichment of resources, providing the hierarchical backbone structure of knowledge bases and allows for assigning types to entities. Taking the example in Fig. 1, while several hypernym relations can be identified, e.g., (*Oxfords*,*Men's shoes*) or (*Derby*, *Men's shoes*), few of them are present in BabelNet and none in DBpedia.

We first propose a knowledge extraction approach for identifying hypernym relations carried out by VES. We implement a learning approach for the following reasons (1) the corpus has many regularities that can emerge with this kind of approach and (2) features of different nature (syntactic, lexical, typographical, dispositional, semantic or distributional) can be combined together. In particular, the choice of a distant learning is motivated by the fact that it is free of manual annotation and that the learning knowledge base (here BabelNet) and the learning text (Wikipedia) are aligned, as recommended by the method. We then evaluate the enrichment rate from an experiment we led on a corpus made of VES extracted from French Wikipedia pages.

This work is part of the SemPedia[2] project aiming at enriching DBpedia for French, by specifying and implementing a set of new Wikipedia extractors dedicated to the hypernym relation. We focus on French because semantic resources targeting this language are scarce. We have already proposed a distant supervised approach and implemented a tool for identifying hypernym relations from disambiguation Wikipedia pages [15]. We propose to adapt this approach in this new context, i.e. identifying hypernym relations from vertical enumerative structure of Wikipedia pages, ensuring that the approach is:

– free of manual annotation;

[1] A hypernym relation link two entities E_1 and E_2 when E_2 (hyponym) is subordinate to E_1 (hypernym). From a lexical point of view, this relation is called "isa".

[2] http://www.irit.fr/Sempedia.

Men's shoes can be categorized by how they are closed:

- Oxfords (also referred as "Balmorals"): the vamp has a V-shaped slit to which the laces are attached; also known as "closed lacing". The word "Oxford" is sometimes used by American clothing companies to market shoes that are not Balmorals, such as Blüchers.
- Derby shoe: the laces are tied to two pieces of leather independently attached to the vamp; also known as "open lacing" and is a step down in dressiness. If the laces are not independently attached to the vamp, the shoe is known as a blucher shoe. This name is, in American English, often used about derbys.
- Monk-straps: a buckle and strap instead of lacing.
- Slip-ons: There are no lacings or fastenings. The popular loafers are part of this category, as well as less popular styles, such as elastic-sided shoes.

Fig. 1. Example of VES (https://en.wikipedia.org/wiki/Shoe)

- language independent, thus may be reused for enriching DBpedia in several languages;
- reproducible on any corpus which contains VES having same discourse properties that those of Wikipedia pages.

The rest of this paper is structured as follows. Section 2 presents the background on enumerative structures and on distant learning. We present then our learning model in Sect. 3. Section 4 describes the experimentation and the evaluation. Section 5 discusses the main related work. Finally, Sect. 6 concludes the paper and discusses future work.

2 Background

In this section, we first describe the main principles of distant learning. We then introduce the discursive properties of enumerative structures we lean on to implement our approach.

2.1 Distant Learning

The distant learning method follows the same principles as the supervised learning ones, except that the annotation for constructing the learning examples is carried out using an external semantic resource. In the context of relation extraction from text, it consists in aligning an external knowledge base to a corpus and in using this alignment to learn relations [6,21]. The learning ground is based on the hypothesis that "if two entities participate in a relation, all sentences that mention these two entities express that relation". Although this hypothesis seems too strong, Riedel *et al.* [26] show that it makes sense when the knowledge base used to annotate the corpus is derived from the corpus itself. Thus, for a pair of entities appearing together within a sentence, a set of features are extracted from the sentence and added to a feature vector for that entity pair.

If the entities are linked in the knowledge base, that entity pair constitutes a positive example, a negative example otherwise.

This approach has been exploited for identifying relations expressed in sentences which are syntactically and semantically correct. Our contribution relies on adapting this approach to different textual structures, especially for those where a part of semantics is carried out by layout. For each type of textual structure, it is then necessary to define a process for building learning examples and to define discriminant features.

2.2 Vertical Enumerative Structures

An enumerative structure (ES) is a textual structure which expresses hierarchical knowledge through different components. According the definition of Ho-Dac *et al.* [12], "it encompasses an *enumerative theme* justifying the union of several elements according to an identity of statut". Different types of enumerative structures exist and different typologies have been proposed.

From a visual point of view, a vertical ES (VES) is expressed using typographic and dispositional markers. More specifically, a VES is composed of (1) a primer (corresponding to a sentence or a phrase) which contains the "enumerative theme" and which introduces (2) a list of items (at least two items) which belong to the same conceptual domain, and (3) possibly of a conclusion. If we consider the example of Fig. 1, "Men's shoes can be categorized by how they are closed:" is the primer, "Oxfords ... such as Blüchers." is an item, *Men's shoes* is the "enumerative theme" and *Oxfords*, *Derby shoes*, *Monk-straps* and *Slip-ons* are entities of the same conceptual domain. This VES has no conclusion.

From a discursive point of view, VES may be classified according to the discourse relations between their components. Before introducing VES properties our approach relies on, we first briefly remind the major principles of the Segmented Discourse Representation Theory (SDRT) [1] which is the discourse theory we used for analyzing VES. A discourse analysis in that context consists in breaking down the text into segments (called discourse units or DU) and in linking adjacent segments with *coordinating* or *subordinating* relations. *Coordinating* relations link entities of the same importance, whereas *subordinating* relations link an entity to an entity of lower importance. Thus, if we consider the primer and items of a VES as DUs (resp. DU_{Primer} and $DU_{Item_j}(j = 1, ..., N)$ if VES is composed of N items), a manual discourse analysis of such a VES allows to state that the primer is linked to the first item with a *subordinating* relation. When all items are linked with *coordinating* relations, we qualify such VES as *paradigmatic* [9] and refer to it as P-VES (Fig. 2(a)).

According again to the SDRT, if DU_{Item_1} is subordinated to DU_{Primer}, hence each DU_{Item_j} coordinated to $DU_{Item_{(j-1)}}$ (j = 2, ..., N), is subordinated to DU_{Primer}. Thereby, N subordinating relations between DU_{Primer} and DU_{Item_j}, (j = 1,...,N), can be inferred (Fig. 2 (b)). In that context and as the *elaboration* relation is a sub-relation of the *subordinating* one, we can also say that each DU_{Item_j} *elaborates* DU_{Primer}. When DU_{Primer} and DU_{Item_j} are broken down into more fine-grained DUs as terms, these discourse relations are kept between

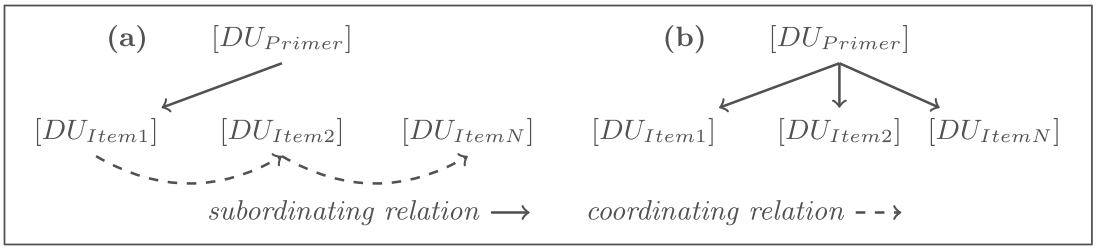


Fig. 2. Discursive representations of P-VES according to the SDRT.

at least one term H in the primer and one term h in the item. From a lexical point of view, these N relations may be specialized in at least N lexical relations $\mathrm{R(H,\ h}_i)_{i=1,\ldots,N}$.

We are interested with P-VES as Wikipedia pages contain many P-VES often expressing definitions and properties of entities. These pages are written according to the guide "The Manual Of Style"[3] which recommends the same grammatical form for all items. An analysis of 100 Wikipedia pages randomly chosen shows that more than 80% of VES respect those instructions and thus are paradigmatic.

We are aware that a P-VES can, however, bear more relations (hierarchical or no hierarchical). The example in Fig. 3 shows that more than one hierarchical relations exist between the primer and the first item which is itself composed of a list (*act of worship* and *sacrifice*, *act of worship* and *libation*, etc.), as well as one no hierarchical (syntagmatic) relation expressed in the last item (*preaching* and *Abrahamic religions*).

The main acts of worship are:

- sacrifice, libation, the offering and education;
- pray (invocation, praise, etc.);
- preaching which plays an important role in Abrahamic religions;

Fig. 3. P-VES containing hierarchical relations and one no hierarchical relation.

3 Proposed Approach

We describe here how the distant learning approach has been adapted for learning hypernym relations from P-VES. We describe in particular the process of building learning examples and the learning model.

[3] http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style.

3.1 Learning Examples Building

The process of building the examples is composed of four main steps, which are detailed in the following.

Step 1: distribution of the primer over the items. On the basis of the discursive properties of P-VES (Sect. 2.2) where each item *elaborates* (in the sense of the *elaboration* discourse relation) the primer, we distribute the primer over each item, giving rise to new textual units called TU. Each *TU* corresponds to the concatenation of the primer and of one item. *N TU* are thus generated if the P-VES is composed of *N* items. For illustrating this step, we consider the P-VES depicted in Fig. 1. Four TU are thus generated:

1. [Men's shoes ... are closed: Oxfords ... such as Blüchers.]
2. [Men's shoes ... are closed: Derby shoe... often used about derbys.]
3. [Men's shoes ... are closed: Monk-straps: ... instead of lacing]
4. [Men's shoes ... are closed: Slip-ons... such as elastic-sided shoes.]

Distributing the primer over the different items does not make us fall back on the classic extraction of relations because we are still confronted with the presence of some typographic markers that replace lexical markers.

Step 2: annotation of terms. In this step, we firstly extract the terms from the external semantic resource (i.e. the terminology provided by the resource, such as synset terms or concept labels). Then, this list is used to annotate the set of generated TUs.

Step 3: building the couples of terms. Several terms may be present in the primer or in an item. Learning examples are then built from couples of terms ($Term_1$,$Term_2$) which respectively belong to the primer and to one of the items. Relying on the *elaboration* relation between the primer and an item, we empirically define the following heuristics for selecting couples of terms:

- $Term_2$ should belong to the first part of the item, i.e. the string starting at index 0 and ending at the next final punctuation (point, line return, etc.) of the item. As Fig. 1, an item may be composed of several sentences.
- Items for which no terms are linked to at least one term of the primer by a hypernym relation, according to the external semantic resource, are left aside. Indeed, this case contradicts our underlying assumption about P-VES that states that each item *elaborates* the primer. This case may nevertheless be explained by a possible incompleteness of the external resource.

For constructing the set of examples, we have generated all the combinations of terms ($Term_1$, $Term_2$) from the primer and retained items. Thus a couple of terms will correspond to a positive example if a hypernym relation between $Term_1$ and $Term_2$ exists in the external semantic resource, to a negative example otherwise. We are aware however that, depending on the coverage of the external semantic resource adopted, negative examples may be false negative ones given the fact that the relation is simply missing in the resource.

Step 4: associating a set of features to couple of terms. Each couple of terms $Term_1$ and $Term_2$ is associated with a set of features coming from the textual units from which they have been extracted. Currently, we focus on lexical features, grammatical features, layout features, and some heuristics inspired by [18], such as the context of $Term_1$ and $Term_2$ (text window). Furthermore, features impact different levels of P-VES: those involving the whole enumerative structure, those involving the example, those involving the primer and those involving the item. Table 1 introduces the set of selected features as used in the experiments described in Sect. 4.

Table 1. Set of learning features (*an enumerative theme is used for organizing the concepts involved into an enumerative structure and is one of the following expressions *list of*, *types of*, *kind of*, etc.).

Scope	Features	Description	Datatype
ES	itemsNumber	number of items present in the VES	integer
Example	lexicalInclusion	lexical inclusion between the terms	boolean
Primer	nbTokens_P	number of tokens in the primer	integer
	lemmaPOSWindow_P	sequence of POS of the window corresponding to 3 tokens preceding $Term_1$, 3 tokens following $Term_1$	string
	lemmaPosTerm1	sequence of POS of all tokens included into $Term_1$	string
	NbTokensBeforeTerm1	number of tokens before $Term_1$	int
	NbTokensAfterTerm1	number of tokens after $Term_1$	int
	capitalizedInitialTerm1	initial of $Term_1$ is capitalized	boolean
	capitalizedTerm1	$Term_1$ is capitalized	boolean
	endsWithColon	primer ends with a colon	boolean
	verbPresence	the primer contains a verbal form	boolean
	theme*	the primer contains an enumerative theme	boolean
	nbTokensTerm1Org	number of tokens between $Term_1$ and the theme	integer
	ordinal	the primer contains a numeral	boolean
	nbTokensTerm1Ord	number of tokens between $Term_1$ and the numeral	integer
Item	nbTokens_I	number of tokens in the item	integer
	nbSentences_I	number of sentences in the item	integer
	lemmaPOSWindow_I	sequence of POS of the window including 3 tokens preceding $Term_2$, 3 tokens following $Term_2$	string
	lemmaPosTerm2	sequence of POS of all tokens included into $Term_2$	string
	NbTokensBeforeTerm2	number of tokens before $Term_2$	int
	NbTokensAfterTerm2	number of tokens after $Term_2$	int
	capitalizedInitialTerm2	initial of $Term_2$ is capitalized	boolean
	capitalizedTerm2	$Term_2$ is capitalized	boolean

3.2 Learning Model

In order to perform a binary classification task (*isA* or *not-isA* classes), we chose the Maximum Entropy classifier (MaxEnt) [3] which is relevant when the conditional independence of the features cannot be assured. This is particularly true in NLP where features are usually words which obviously are not independent in their use (they are bound by syntactic and semantic rules). Furthermore, MaxEnt allows the management of a high number of features. It relies on the maximum entropy principle. Hence, it requires to define a set of constraints for each observation and to choose the distribution which maximizes the entropy

while remaining consistent with the whole set of constraints [14]. In this context of optimisation under constraints, it is mathematically proved that a unique solution exists and that an iterative algorithm converges towards this solution [25]. The classical formula of MaxEnt is the following:

$$P(y|x) = \frac{1}{Z} \exp\left(\sum_i w_i f_i(x, y) \right)$$

where $P(y|x)$ gives the probability that the individual x (here a relation) belongs to the class y (here *isA* or *not-isA* classes). Each individual is encoded as a feature vector. The function f_i is a function called *feature* which determines the constraints of the model. The weights w_i associated to each feature account for the probability to belong to a class. Z is a normalization constant which ensures that the sum of probabilities of one individual is equal to 1.

To estimate the parameter values $\hat{w}$, we use the likelihood function that aims at determining the best estimators:

$$\hat{w} = argmax \sum_j log(P(y_j|x_j))$$

where the (x_j, y_j) belongs to the set of training data. We used the OpenNLP (version 1.5.0) implementation of the MaxEnt algorithm[4].

4 Experiments

Our experiments were carried out on the French Wikipedia corpus aiming at enriching French DBpedia, using BabelNet as external semantic resource. These choices are motivated by the fact that (1) semantic resources targeting French language are scarce (French DBpedia is about 20,000 times poorer than DBpedia in English); (2) BabelNet [23] is a multilingual network of concepts and named entities that results from the automatic integration of various background knowledge resources (WordNet, Open Multilingual WordNet, Wikipedia, GeoNames, WoNef, etc.); and the learning knowledge base (BabelNet) and the learning text (Wikipedia) are aligned, as recommended by the method. While in BabelNet some mappings (Wikipedia-WordNet) have been manually checked, YAGO [29] assures a better accuracy of the whole knowledge base. However, we choose BabelNet due to its better coverage of French. It consists of about 14 million entries, including concepts and named entities. Using BabelNet publicly available resource allows for a reproducible approach. We used BabelNet 3.7 version.

4.1 Corpus

We built a corpus from a set of enumerative structures extracted from the 2016 dump of French Wikipedia pages. We have used the WikiExtractor tool[5] for

[4] http://opennlp.apache.org/.
[5] https://github.com/attardi/wikiextractor.

extracting plain text of vertical lists based on HTML tags of lists. We then preprocessed the extracted structures by (1) removing the (multiple) malformed lists; (2) reducing the primer to its last sentence, when the WikiExtractor has extracted a whole paragraph as a primer; (3) reformatting each enumerative structure according to the XML schema we defined; and (5) annotating the corpus with BabelNet label concepts and processing the corpus using Tokenizer, SentenceSplitter, TreeTagger, Gazetteer tools available in the GATE system[6]. This resulted in a corpus of 75446 annotated enumerative structures.

4.2 Learning Examples

From the corpus, 134170 examples were built (2766 positive examples and 131404 negative examples), as the method described in Sect. 3.1:

- an example is positive if $Term_1$ (present in the primer) and $Term_2$ (present in an item) are linked by a direct hypernym relation in BabelNet. We could observe that, given the polysemous nature of terms, considering a high path in the network between them introduce noise. We have then restricted to length 1 the path for classifying the example as positive;
- an example is negative if no path of length lower than 3 exists between $Term_1$ and $Term_2$ in BabelNet. This relaxes the assumption in Sect. 3.1, where examples are assumed to be negative if the relation is simply missing in the resource. From an empirical analysis, we fix to 3 the path length. We could observe that, even if two terms are linked in the resource (with a path length higher than 3), this link may not reflect a hierarchical relation given the polysemous nature of the terms (terms loosely related).

We thus built a training set of 3688 examples (1844 positive and 1844 negative examples) and a test set made of 1844 examples (922 positive and 922 negative examples). Examples have been randomly chosen among the initial sets of positive and negative examples.

4.3 Evaluation Setting

We evaluated our approach on the test set and on 2 reference corpora. These reference corpora have been used by [8] for evaluating their supervised learning approach intended to also identify hypernym relations from enumerative structures. To the best of our knowledge, this is the only corpus of same nature available for comparison. Each reference corpus concerns a set of Wikipedia pages having the same topic, respectively *Computer science* and *Transport*. They have been annotated with terms obtained from both YaTeA and Acabit term extractors. In that work, the results have been reported only in terms of precision on the top 500 hypernym relations extracted. Here, we have used these reference corpora and the reported results as baseline.

[6] https://gate.ac.uk/sale/tao/splitch6.html#chap:annie.

The reference sets correspond to examples built from these annotated reference corpora. We have produced several learning models, varying the different features, and with different sizes of sentence. A linear regression analysis of features gives features listed in Table 1 as discriminant. We keep the model (we named DLM for Distant Learning Model) taking into account these features.

4.4 Results and Discussion

Table 2 presents the results of our approach considering the test set, in terms of precision, recall, F-measure and accuracy. Table 3 presents those considering the reference corpus, in terms of precision. We can observe good values of precision and recall on the test set, while observing varying results in terms of precision on the two reference sets. The low performance of our approach on the *Transport* corpus can be explained by two main reasons. First, this corpus contains several contextual spatial relations, which are expressed using nested VES, where the context is expressed in the primer of one of these nested VES. However, our approach takes the VES independently of each other and hence can not correctly deal with this contextual parameter. Second, these VES are generally composed of numerous items. For the *Computer* corpus our approach outperforms the baseline. Overall, we obtain a precision up to 0.62.

Table 2. Results for the test set for all features of Table 1.

	Precision	Recall	F-measure	Accuracy
DLM on the test set	0.73	0.83	0.78	0.76

Table 3. Results for the reference set for all features of Table 1.

	Precision
DLM on Computer reference set	0.73
DLM on Transport reference set	0.51
Baseline on Computer reference set	0.6
Baseline on Transport reference set	0.6

We could identify some sources of noise in the learning process. First, we could observe that the external resource used here is not exhaustive, which may lead to the generation of false negative examples. That goes against the hypothesis of distant learning approach. Second, false positive examples are introduced due to the fact that the term ambiguity is propagated when exploring the network. Third, the knowledge may be expressed in a different way according to the language. In fact, we can observe cycles in the French network. For instance, the

cycle “microprocessor” is a “microprocessor” in the French network does not exist in the English one.

Besides that, our approach is able to correctly identify the hypernym relations between textual entities (primer, item) that are not contiguous in the text, such as the set of hypernym relations in Fig. 1. These kinds of relations can not be in fact correctly treated by the classical NLP parsers. Furthermore, we can observe that the model is able to correctly identify the cases of head modifiers, and identifies hypernym relations such as (*Ministères*, *Ministère de la Sécurité publique*), (*Ministères*, *Ministère de la Supervision*), (*Ministères*, *Ministère de la Justice*), as from the VES in Fig. 4.

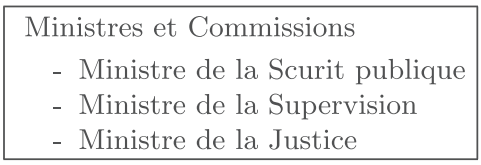


Fig. 4. Hypernym relations identified from head modifiers.

4.5 DBpedia Enrichment

We have evaluated how much our approach could enrich DBpedia with the extracted hypernym relations. These relations contribute to enrich DBpedia in two ways: terms participating in the relations and relations themselves. To do so, we checked the presence/absence in DBpedia of them. The list of checked relations comes from the set of 307 true positive (TP) relations classified by our training model (with a confidence ≥ 0.5) on the reference corpus (168 TP from the *Computer* corpus and 139 TP from the *Transport* corpus).

First of all, we checked how many of the annotated terms are present in DBpedia. For that, we followed two strategies: (i) using a SPARQL query with an exact match between the relation terms and labels of DBpedia resources and (ii) using the DBpedia Spotlight service[7], a tool for automatically annotating mentions of DBpedia resources in texts [7]. We have used the Docker available for French[8]. With the first strategy, for the *Computer* corpus, we found 20 (out of 192) terms in DBpedia, against 9 (out of 168) terms from the *Transport* corpus. Using DBpedia Spotlight, with a confidence of 1 and a support of 20, we found only 3 out of 192 terms from the *Computer* corpus against 6 out of 168 terms from the *Transport* corpus. All of these terms referring to named entities. With a confidence of 0.6, we found 40 terms from the *Computer* corpus and 49 from the *Transport* corpus. However, wrong correspondences have been

[7] https://www.DBpedia-spotlight.org/.
[8] https://github.com/DBpedia-spotlight/spotlight-docker/tree/master/nightly-build/french.

identified lowering the confidence, as somehow expected. This shows that most of the annotated terms from the reference corpus can in fact enrich the resource.

With respect to the 307 TP relations annotated by our model, Table 4 shows the number of relations existing in DBpedia with respect to the presence of their terms in the resource. We can observe that, although some terms participating in the identified relations are present in the resource, only 4 of them participate in the same relation (Spotlight conf = 0.6 in Table 4). Looking at the number of TP relations present in DBpedia, 99% of them are not present in DBpedia. These results confirm that the Wikipedia pages, which are under-exploited by Wikipedia extractors, provide rich hypernym relations other than those found in structured elements (infoboxes, categories, etc.).

Table 4. Presence of relations and their corresponding terms in DBpedia.

	2 terms in DBpedia	Only $Term_1$	Only $Term_2$	None them	Number of present relations
Computer (exact match)	0	17	2	149	0
Transport (exact match)	0	5	3	131	0
Computer (Spotlight conf = 1)	0	6	0	162	0
Transport (Spotlight conf = 1)	0	3	0	136	0
Computer (Spotlight conf = 0.6)	1	35	3	129	0
Transport (Spotlight conf = 0.6)	3	41	2	93	2

5 Related Work

Our approach is related to three main fields of study, whose main related works are discussed in the following.

Enumerative Structures. Firstly, concerning the works on ES, we can mention typologies such as the one proposed by Vergez *et al.* [31], where the items can be present or not in the primer (one-step vs. two-step), or that of Ho-Dac *et al.* [12] where ESs have been classified according to their level of granularity (intra-paragraphic vs. multi-paragraphic). Concerning VES (particularly studied in the context of text generation), Hovy and Arens [13] distinguish the list of items (set of elements of same level) from the enumerated list (for which the order of items is important), while Luc [19] proposes a typology that opposes parallel ES (paradigmatic, visually homogeneous and isolated) to non-parallel ES. This latter is based on the composition of the rhetorical model of Rhetorical Structure Theory (RST) [20] and Textual Architecture Model (TAM) [32]. Drawing inspiration from these works, we also proposed a typology of ES in [9] (we refer to this topology in Sect. 2.2) relying on its discursive properties. These are the discursive properties of paradigmatic VES we exploit in this work.

Hypernym Relation Extraction. Numerous studies have been done in this field and a relevant short overview of them can be found in [33]. The pioneering

work of the linguistic methods is that of Hearst [11] which defined a set of lexico-syntactic patterns specific to the hypernym relation for English. This work has been adapted and extended to improve recall for instance with the concept of "star-pattern" [24], or by progressively integrating learning techniques. Snow *et al.* [28] and Bunescu *et al.* [5] apply supervised learning methods to a set of manually annotated examples. While the cost of manual annotation is the main limit of supervised learning, distant learning method consists in building the set of examples using an external resource to automatically annotate the learning examples. For instance, Mintz *et al.* [21] use Freebase as external resource in their distant approach for identifying around 102 different relations. They implement a multi-class optimized logistic classifier using L-BFGS with Gaussian regularization. Another way to avoid manual annotation is the bootstrapping which uses a selection of patterns to construct the set of examples [4]. Some of these works are based on distributional analyses [17].

Enriching Semantic Resources. With respect to the exploitation of Wikipedia for the construction and enrichment of semantic resources, several extractors have been developed to analyze each type of structured data in Wikipedia. Morsey *et al.* [22] developed 19 of such extractors that produce a formal representation of entities and relations identified within various structural elements from Wikipedia: abstracts, images, infoboxes, etc. Other works have targeted specific relations, mainly hypernym relations. For example, Suchanek *et al.* [29] used the User Generated Categories (UGCs) hierarchy of Wikipedia to build hypernym relations in the Yago knowledge base. Kazama and Torisawa [16] exploited the abstract part of the pages, whereas Sumida and Torisawa [30] extracted knowledge from the menu items. Recent works proposed the automatic creation of MultiWiBi [10], an integrated bitaxonomy of Wikipedia pages and categories in multiple languages. Still, extracted relations from the text in Wikipedia pages have been little used to feed DBpedia [27]. Hence, most of the knowledge from these pages remains unexploited.

Here, we target the extraction of hypernym relations from paradigmatic VES (P-VES), which are under-exploited in the literature aiming at enriching existing semantic resources. A previous approach with the same objectives, based on supervised learning and which exploit semantic properties of a P-VES when considering it as a whole semantic unit, have been proposed by [8]. In that context, a precision of about 60% has been obtained. Given the cost-intensive manual annotation of examples, the expertise required for this task, and in order to improve these results, we adopt a different approach based on a distant supervised learning algorithm, and for which P-VES are no longer semantically considered as a whole unit, but are split into N independent textual units (a textual unit is then composed of the primer and one item) if N is the number of items which compose the P-VES. This approach can be carried out on any corpus presenting structural and/or linguistic regularities, such as web documents, and it is language-independent. Indeed, the proposed approach relies on a multilingual resource that can be used for annotating a corpus and on shallow learning features whose extraction does not depend on deep language analyzers.

6 Conclusion and Future Work

This paper has proposed a knowledge extraction approach that exploits vertical enumerative structure, which are frequent in corpora and are rich sources of hypernym relations. They are however under-exploited by knowledge approaches aiming at enriching semantic resources. We applied a distant learning approach for extracting hypernym relations from vertical enumerative structures expressed in French Wikipedia pages, aiming at enriching the French DBpedia. The aim was at evaluating how hypernym relation extraction can take advantage of enumerative structures for enriching existing resources that have been constructed from the same basis. In that sense, we observed that 99% of the extracted relations could be used for enriching the French DBpedia.

As perspectives, we plan to extend our learning features with additional features such as semantic and distributional features and to deal with the disambiguation of terms, as well as to combine different external resources. We intend as well to apply term extractors in order to identify terms before identifying potential relations that can be used for enriching an existing semantic resource. Finally, we plan to exploit different ontology matching methods in order to integrate the extracted relations into the French DBpedia.

References

1. Asher, N.: Reference to Abstract Objects in Discourse: A Philosophical Semantics for Natural Language Metaphysics. SLAP, vol. 50. Kluwer, Dordrecht (1993)
2. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: dbpedia
3. Berger, A.L., Pietra, V.J.D., Pietra, S.A.D.: A maximum entropy approach to natural language processing. Comput. Linguist. **22**(1), 39–71 (1996)
4. Brin, S.: Extracting patterns and relations from the World Wide Web. In: Atzeni, P., Mendelzon, A., Mecca, G. (eds.) WebDB 1998. LNCS, vol. 1590, pp. 172–183. Springer, Heidelberg (1999). https://doi.org/10.1007/10704656_11
5. Bunescu, R.C., Mooney, R.J.: A shortest path dependency kernel for relation extraction. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 724–731 (2005)
6. Bunescu, R.C., Mooney, R.J.: Learning to extract relations from the web using minimal supervision. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007), Prague, Czech Republic, June 2007
7. Daiber, J., Jakob, M., Hokamp, C., Mendes, P.N.: Improving efficiency and accuracy in multilingual entity extraction. In: Proceedings of the 9th International Conference on Semantic Systems (I-Semantics) (2013)
8. Fauconnier, J.P., Kamel, M.: Discovering hypernymy relations using text layout. In: Joint Conference on Lexical and Computational Semantics, Denver, Colorado, pp. 249–258. ACL (2015)
9. Fauconnier, J.-P., Kamel, M., Rothenburger, B.: Une typologie multidimensionnelle des structures énumératives pour l'identification des relations termino-ontologiques. In: Conférence Internationale sur la Terminologie et l'Intelligence Artificielle - TIA 2013, pp. 137–144, Paris, France, October 2013
10. Flati, T., Vannella, D., Pasini, T., Navigli, R.: MultiWiBi: the multilingual Wikipedia bitaxonomy project. Artif. Intell. **241**, 66–102 (2016). (Complete)

11. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th Conference on Computational Linguistics, pp. 539–545. Association for Computational Linguistics (1992)
12. Ho-Dac, L.-M., Péry-Woodley, M.-P., Tanguy, L.: Anatomie des Structures Énumératives. In: Traitement Automatique des Langues Naturelles, Montréal, Canada (2010)
13. Hovy, E., Arens, Y.: Readings in intelligent user interfaces. In: Automatic Generation of Formatted Text, pp. 256–262. Morgan Kaufmann Publishers (1998)
14. Jaynes, E.: Information theory and statistical mechanics. Phys. Rev. **106**(4), 620 (1957)
15. Kamel, M., Trojahn, C., Ghamnia, A., Aussenac-Gilles, N., Fabre, C.: A distant learning approach for extracting hypernym relations from Wikipedia disambiguation pages. In: International Conference on Knowledge Based and Intelligent Information and Engineering Systems, 6–8 September 2017, France (2017)
16. Kazama, J., Torisawa, K.: Exploiting Wikipedia as external knowledge for named entity recognition. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 698–707 (2007)
17. Lenci, A., Benotto, G.: Identifying hypernyms in distributional semantic spaces. In: Proceedings of the First Joint Conference on Lexical and Computational Semantics, pp. 75–79. Association for Computational Linguistics (2012)
18. Lin, Y., Shen, S., Liu, Z., Luan, H., Sun, M.: Neural relation extraction with selective attention over instances. In: ACL (2016)
19. Luc, C.: Représentation et composition des structures visuelles et rhétoriques du textes. Approche pour la génération de textes formatés. Ph.D. thesis (2000)
20. Mann, W.C., Thompson, S.A.: Rhetorical structure theory: toward a functional theory of text organization. Text **8**(3), 243–281 (1988)
21. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pp. 1003–1011 (2009)
22. Morsey, M., Lehmann, J., Auer, S., Stadler, C., Hellmann, S.: DBpedia and the live extraction of structured data from Wikipedia. Program Electron. Libr. Inf. Syst. **46**, 27 (2012)
23. Navigli, R., Ponzetto, S.P.: BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif. Intell. **193**, 217–250 (2012)
24. Navigli, R., Velardi, P.: Learning word-class lattices for definition and hypernym extraction. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL 2010, Stroudsburg, PA, USA, pp. 1318–1327. Association for Computational Linguistics (2010)
25. Ratnaparkhi, A.: Maximum entropy models for natural language ambiguity resolution. Ph.D. thesis, University of Pennsylvania (1998)
26. Riedel, S., Yao, L., McCallum, A.: Modeling relations and their mentions without labeled text. In: Balcázar, J.L., Bonchi, F., Gionis, A., Sebag, M. (eds.) ECML PKDD 2010. LNCS (LNAI), vol. 6323, pp. 148–163. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15939-8_10
27. Rodriguez-Ferreira, T., Rabadan, A., Hervas, R., Diaz, A.: Improving information extraction from Wikipedia texts using basic English. In: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC) (2016)

28. Snow, R., Jurafsky, D., Ng, A.Y.: Learning syntactic patterns for automatic hypernym discovery. In: Advances in Neural Information Processing Systems 17 (2004)
29. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge unifying WordNet and Wikipedia. In: Proceedings of the 16th International Conference on World Wide Web, WWW 2007, pp. 697–706 (2007)
30. Sumida, A., Torisawa, K.: Hacking wikipedia for hyponymy relation acquisition. IJCNLP **8**, 883–888 (2008)
31. Vergez-Couret, M., Prevot, L., Bras, M.: Interleaved discourse, the case of two-step enumerative structures. In: Proceedings of Contraints In Discourse III, Postdam, pp. 85–94 (2008)
32. Virbel, J.: Structured Documents, pp. 161–180. Cambridge University Press, New York (1989)
33. Wang, C., He, X., Zhou, A.: A short survey on taxonomy learning from text corpora: issues, resources and recent advances. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1190–1203 (2017)

7 Conclusion

7.1 Synthèse des contributions

L'objectif de la thèse était le développement de méthodes d'extraction de connaissances pour l'acquisition de relations hyperonymiques permettant d'enrichir les bases de connaissances pour le français, qui sont aujourd'hui nettement moins riches que leurs équivalents en anglais. Cet objectif, qui a donné lieu à de nombreux travaux, reste un enjeu aujourd'hui dans la mesure où il s'agit d'extraire une information de nature conceptuelle à partir de textes caractérisés par la diversité des formulations, la redondance, l'ambiguïté. Ceci explique que la plupart des travaux visant l'enrichissement des bases de connaissances aient eu tendance à tirer seulement parti de la part structurée d'une base textuelle telle que Wikipedia, laissant inexploitée la partie textuelle.

Nous avons présenté une série d'expériences exploitant le corpus Wikipedia dans sa dimension textuelle et combinant des approches par apprentissage fondées sur le principe de la supervision distante, et des méthodes linguistiques à base de patrons, pour extraire de ce corpus des mentions d'entités ou de classes en relation d'hyperonymie. Ce travail porte sur les pages en langue française de Wikipedia, dans le but de mieux exploiter le langage naturel sur ces pages et d'enrichir la ressource sémantique DBPedia en français. Ce type d'étude est rare pour la langue française.

L'objectif de ces expérimentations était multiple :

— optimiser l'utilisation de méthodes existantes d'extraction de relations sémantiques. Dans le cas des méthodes par patrons, nous avons combiné des ressources complémentaires existant pour le français, et testé l'apport de patrons spécifiques dans le cas des pages de désambiguïsation. Dans le cas des méthodes par apprentissage, nous avons mené différentes expériences pour définir la meilleure combinaison de traits ;
— définir les conditions optimales de leur utilisation, afin de choisir la meilleure méthode en fonction de la nature du matériau textuel disponible ;
— enrichir les méthodes existantes en tirant parti du maximum d'information disponible dans les corpus textuels, combinant ainsi le contenu textuel brut et des indices issus de la mise en forme et de la structure du texte, comme l'exploitation des structures énumératives.
— démontré l'apport des informations extraites des textes pour enrichir des bases de connaissances en français, et plus précisément DBPedia en français.

Chacun de ces objectifs a donné lieu à des expérimentations dont les résultats permettront de guider le choix entre les méthodes disponibles. Ce travail a montré la difficulté qu'il y a à mettre au point des approches efficaces sur un corpus donné, quelle que soit la technique. Il a confirmé l'intérêt de diversifier et multiplier les méthodes pour extraire des relations de manière à exploiter pour chacune une des formes possibles des textes et de l'expression de ces relations. Il adonné lieu a quatre publications présentées lors de conférences nationales et internationales.

7.2 Perspectives

Nous avons commencé à faire des expériences d'enrichissement des bases de connaissances (BabelNet et DBPedia) et montré l'apport effectif de l'extraction de relations entre concepts à partir des textes. Par exemple, on a montré que 99% de l'information accessible à partir des structures énumératives présentes sur des pages de Wikipédia pouvait être utilisée dans cet objectif. Néanmoins, la jonction entre la tâche d'extraction et la tâche d'enrichissement n'a pas été réalisée de façon pleinement opérationnelle dans la thèse, en particulier pour gérer de façon efficace les problèmes d'ambiguïté termes / concept. Sa mise en œuvre est donc le prolongement le plus immédiat de ce travail. Une aide interactive est indispensable pour faire intervenir des spécialistes des connaissances modélisées pour valiser les propositions que font les algorithmes d'extraction de triplets relation(concept1, concept2). Mais cet application interactive doit aussi faciliter l'interprétation d'un ensemble de triplets extraits afin de décider comment les intégrer dans la ressource existante. Ce choix porte autant sur la manière dont les concepts repérés dans le triplets peuvent être rattachés à des concepts déjà existant dans la ressource ou donner lieu à de nouveaux concepts, qu'à la manière de représenter la relation. Dans le cas d'une relation d'hyperonymie entre termes, qui est la relation sur laquelle nous nous sommes focalisés, l'ajout de

triplets se traduit par l'ajout de relations 'sous-classes de' entre classes de la hiérarchie, ou par des relations de typage entre des entités et des classes. Mais certaines relations ne peuvent pas être ajoutées aussi simplement, et vont pratiquement donner lieu à la représentation de plusieurs relations prar exemple, en utilisant des concepts intermédiaires.

D'autres expériences devront être également menées en amont des extracteurs sémantiques qui ont été développés : la phase d'extraction des termes doit être améliorée et complétée, de manière à étendre l'ensemble des concepts accessibles, et en particulier extraire les unités polylexicales. L'approche par supervision distante gagnerait certainement à être enrichie par l'apport d'autres bases de connaissances externes. Par ailleurs, les possibilités offertes par l'information distributionnelle ont certes été testées pour enrichir les traits des classifieurs, mais leur intégration devra être expérimentée de façon plus systématique afin d'améliorer les performances des extracteurs que nous avons développés.

De plus, si la relation d'hyperonymie, qui a été au centre de nos travaux, constitue le mode de structuration privilégié des bases de connaissances, d'autres relations devront être prises en compte pour enrichir ces ressources sémantiques, qu'il s'agisse d'autres relations lexicales (comme la relation partie-tout) ou de relations conceptuelles diverses (localisation, possession, but, causalité, ou encore rôles, etc.).

Enfin, rappelons que le domaine de l'extraction de relations est un domaine très actif depuis quelques années, avec de nouveaux résultats et de nouveaux algorithmes améliorant les performances chaque 6 mois. Plusieurs jeux de données avec des annotations et des résultats attendus sont disponibles pour la langue anglaise, intégrant soit des corpus généraux et des relations générales (campagnes SemEval), soit des domaines et corpus spécialisés comme les texte biomédicaaux (d'autres campagnes SemEval et la campagne BioNLP). Nous avons peu utilisés ces travaux récents malgré leur performances meilleures, qui sont en grande majorité à base de réseaux de neurones, faute de temps. En effet, nous avons donné priorité à des travaux intégrant une compréhension linguistique des phénomènes. Toutefois, il est indispensable que nous les adaptions maintenant à la langue française et à des sous-ensemble du corpus Wikipedia pour atteindre de meilleures taux de performance, nous intéresser à toutes sortes de relations et enrichir encore la panoplie des techniques utilisées.

Pour terminer, cette étude comporte des enjeux plus larges. D'une part, notre travail permet de rendre accessibles sous forme structurée (plus précisément sous forme de graphes de connaissances) des informations présentes dans des collections de textes. Une fois déposées au sein d'entrepôts de données réutilisables, il est plus facile de développer des applications intégrant ces connaissances. Ces graphes peuvent servir ensuite à enrichir ou mieux interpréter d'autres types de données (des images satellites, des données scientifiques ou de réseaux sociaux) au sein d'applications aussi variées que la recherche d'informations, le guidage routier ou la détection de changements à la surface de la Terre. Au moment d'associer ces connaissances à des ressources textuelles peu structurées, il est important qu'elles soient étiquetées et décrites par des termes en français, et qu'on sache trouver dans des textes quelles relations sont mentionnées. C'est ce qui a justifié notre étude et ses perspectives.

Bibliographie

Khurshid Ahmad and Patrick Holmes-Higgin. System quirk : A unified approach to text and terminology. In *TermNet symposium*, 1995.

Luis Espinosa Anke. Towards definition extraction using conditional random fields. In *RANLP*, pages 63–70, 2013.

Patrick Arnold and Erhard Rahm. Automatic extraction of semantic relations from wikipedia. *International Journal on Artificial Intelligence Tools*, 24(2), 2015. doi : 10.1142/S0218213015400102. URL `https://doi.org/10.1142/S0218213015400102`.

J. Arpirez, Fernandez-Lopez M. Corcho, O., and Gomez-Perez A. Webode in a nutshell. *AI Magazine*, 24(3) :37–48, 2003.

Nicholas Asher. *Reference to abstract objects in discourse*, volume 50 of *Studies in linguistics and philosophy*. Kluwer, 1993. ISBN 978-0-7923-2242-9.

Sophie Aubin and Thierry Hamon. Improving term extraction with terminological resources. In *Advances in Natural Language Processing*, pages 380–387. Springer, 2006.

Alain Auger and Caroline Barrière. Pattern-based approaches to semantic relation extraction : A state-of-the-art. *Terminology*, 14(1) :1–19, 2008.

N. Aussenac-Gilles, S. Despres, and S. Szulman. The terminae method and platform for ontology engineering from texts. In P. Cimiano P. Buitelaar, editor, *Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text*, pages 199–223. IOS Press, 2008.

Nguyen Bach and Sameer Badaskar. A review of relation extraction. *Literature review for Language and Statistics II*, 2, 2007.

Caroline Barrière. Building a concept hierarchy from corpus analysis. *Terminology*, 10-2 :241–263, 2004.

Caroline Barrière and Agbado A. Terminoweb : a software environment for term study in rich contexts. In *International Conference on Terminology, Standardization and Technology Transfert (TSTT 2006)*, pages 103–113, Beijing (China), 2006.

Adam L Berger, Vincent J Della Pietra, and Stephen A Della Pietra. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1) :39–71, 1996.

Matthew Berland and Eugene Charniak. Finding parts in very large corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 57–64. Association for Computational Linguistics, 1999.

Andrée Borillo. Exploration automatisée de textes de spécialité : repérage et identification de la relation lexicale d'hyperonymie. *Linx*, 34(1) :113–124, 1996.

D. Bourigault, N. Aussenac-Gilles, and J. Charlet. Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle (RIA)*, 18(1) :87–110, 2004.

Didier Bourigault. Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In Atala, editor, *TALN 2002*, pages 75–84, Nancy, France, 2002.

Ronald J Brachman and James G Schmolze. An overview of the kl-one knowledge representation system. In *Readings in artificial intelligence and databases*, pages 207–230. Elsevier, 1989.

Paul Buitelaar and Philipp Cimiano, editors. *Ontology Learning and Population : Bridging the Gap between Text and Knowledge*, volume 167 of *Frontiers in Artificial Intelligence and Applications Series*. IOS Press, 2008.

Paul Buitelaar, Daniel Olejnik, and Michael Sintek. Ontolt : A protégé plug-in for ontology extraction from text. In *In Proceedings of the International Semantic Web Conference (ISWC), demo session*, pages 31–44, 2003.

Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini. Ontology learning from text : An overview. In *Ontology Learning from Text : Methods, Evaluation and Applications*, pages 3–12. IOS Press, 2005.

Paul Buitelaar, Philipp Cimiano, Stefania Racioppa, and Melanie Siegel. Ontology-based information extraction with SOBA. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk, and Daniel Tapias, editors, *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 22-28, 2006.*, pages 2321–2324. European Language Resources Association (ELRA), 2006. URL `http://www.lrec-conf.org/proceedings/lrec2006/summaries/93.html`.

Razvan C Bunescu and Raymond J Mooney. A shortest path dependency kernel for relation extraction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 724–731. Association for Computational Linguistics, 2005.

Silvana Castano, Alfio Ferrara, Stefano Montanelli, and Davide Lorusso. Instance matching for ontology population. In Salvatore Gaglio, Ignazio Infantino, and Domenico Saccà, editors, *Proceedings of the Sixteenth Italian Symposium on Advanced Database Systems, SEBD 2008, 22-25 June 2008, Mondello, PA, Italy*, pages 121–132, 2008.

Jinxiu Chen, Donghong Ji, Chew Lim Tan, and Zhengyu Niu. Unsupervised feature selection for relation extraction. In *Proceedings of IJCNLP*, 2005.

Philipp Cimiano and Johanna Völker. Text2onto - a framework for ontology learning and data-driven change discovery. In Andres Montoyo, Rafael Munoz, and Elisabeth Metais, editors, *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB 2005)*, volume 3513 of *Lecture Notes in Computer Science*, pages pp. 227–238, Alicante, Spain, June 2005. Springer. URL `http://code.google.com/p/text2onto/`.

Philipp Cimiano and Johanna Völker. Text2onto. In Andrés Montoyo, Rafael Muńoz, and Elisabeth Métais, editors, *Natural Language Processing and Information Systems*, pages 227–238, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-32110-1.

Anne Condamines and Marie-Paule Jacques. Le repérage de l'hyperonymie par un faisceau d'indices : mise en question de la notion de " marqueur ". In *Semaine de la connaissance*, 30062006, France, June 2006. URL `https://halshs.archives-ouvertes.fr/halshs-01321031`.

Aron Culotta, Andrew McCallum, and Jonathan Betz. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 296–303, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. doi : 10.3115/1220835.1220873. URL `https://doi.org/10.3115/1220835.1220873`.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. Gate : an architecture for development of robust hlt applications. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 168–175. Association for Computational Linguistics, 2002.

Tablan V. Roberts A. Cunningham, H. and K. Bontcheva. Getting more out of biomedical documents with gate's full lifecycle open source text analytics. *PLoS computational biology*, 9(2), 2013. URL `doi:10.1371/journal.pcbi.1002854`.

Koen Deschacht and Marie-Francine Moens. Efficient hierarchical entity classifier using conditional random fields. In *Proceedings of the 2nd Workshop on Ontology Learning and Population*, pages 33–40, 2006.

Xin Luna Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge vault : A web-scale approach to probabilistic knowledge fusion. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'14, New York, NY, USA - August 24 - 27, 2014*, pages 601–610, 2014. URL `http://www.cs.cmu.edu/~nlao/publication/2014.kdd.pdf`.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. Open information extraction from the web. *Commun. ACM*, 51(12) :68–74, December 2008. ISSN 0001-0782. doi : 10.1145/1409360.1409378. URL http://doi.acm.org/10.1145/1409360.1409378.

Cécile Fabre and Alessandro Lenci. Distributional Semantics Today Introduction to the special issue. *Traitement Automatique des Langues*, 56(2) :7–20, 2015. URL https://hal.archives-ouvertes.fr/hal-01259695.

Chimène Fankam, Stéphane Jean, Guy Pierra, and Ladjel Bellatreche. Enrichissement de l'architecture ansi/sparc pour expliciter la sémantique des données : une approche fondée sur les ontologies. *Revue des Nouvelles Technologies de l'Information*, 2ème Conférence Francophone sur les Architectures Logicielles (CAL 2008), 3-7 Mars 2008, Montréal, Québec, Canada, RNTI-L-2 : 47–62, 2008.

Jean-Philippe Fauconnier. *Acquisition of semantic relations from layout elements*. Theses, Université de Toulouse, January 2016. URL https://tel.archives-ouvertes.fr/tel-01324765.

Jean-Philippe Fauconnier and Mouna Kamel. Discovering hypernymy relations using text layout. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 249–258, Denver, Colorado, June 2015. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/S15-1030.

Ralph Grishman. Information extraction : Techniques and challenges. In Maria Teresa Pazienza, editor, *Information Extraction A Multidisciplinary Approach to an Emerging Information Technology*, pages 10–27, Berlin, Heidelberg, 1997. Springer Berlin Heidelberg. ISBN 978-3-540-69548-6.

Thomas R. Gruber. A translation approach to portable ontology specifications. *KNOWLEDGE ACQUISITION*, 5 :199–220, 1993.

Nicola Guarino. Formal ontology and information systems. In *Proceedings of FOIS'98, Trento, Italy, 6-8 June 1998*, pages 3–15. IOS Press, 1998.

Braden Hancock, Paroma Varma, Stephanie Wang, Martin Bringmann, Percy Liang, and Christopher Ré. Training classifiers with natural language explanations. In *ACL 2018*, volume abs/1805.03818, 2018. URL http://arxiv.org/abs/1805.03818.

Zellig S. Harris. *A Theory of Language and Information : A Mathematical Approach*. Clarendon Press, Oxford, 1991.

Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, COLING '92, pages 539–545, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics. doi : 10.3115/992133.992154. URL http://dx.doi.org/10.3115/992133.992154.

Marie-Paule Jacques and Nathalie Aussenac-Gilles. Variabilité des performances des outils de TAL et genre textuel. Cas des patrons lexico-syntaxiques. *Traitement Automatique des Langues, Non Thématique*, 47(1) :(en ligne), 2006. URL http://www.atala.org/Variabilite-des-performances-des.

Jun'ichi Kazama and Kentaro Torisawa. Exploiting wikipedia as external knowledge for named entity recognition. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 698–707, 2007.

Sanghee Kim, Harith Alani, Wendy Hall, Paul H. Lewis, David E. Millard, Nigel R. Shadbolt, and Mark J. Weal. Artequakt : Generating tailored biographies with automatically annotated fragments from the web. In Siegfried Handschuh, Nigel Collier, Rose Dieng, and Steffen Staab, editors, *Proceedings of the ECAI 2002 Workshop on Semantic Authoring, Annotation & Knowledge Markup, SAAKM@ECAI 2002, Lyon, France, July 22-26, 2002*, volume 100 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2002. URL http://ceur-ws.org/Vol-100/Sanghee_Kim-et-al.pdf.

Georges Kleiber and Irène Tamba. L'hyponymie revisitée : inclusion et hiérarchie. *Langages*, 98 : 7–32, 1990.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16(4) :359–389, 2010.

Shantanu Kumar. A survey of deep learning methods for relation extraction. *CoRR*, abs/1705.03645, 2017.

Luce Lefeuvre. *Analyse des marqueurs de relations conceptuelles en corpus spécialisé : recensement, évaluation et caractérisation en fonction du domaine et du genre textuel*. PhD thesis, Thèse de doctorat en Sciences du langage, Université Toulouse 2, 2017. URL `http://www.theses.fr/2017TOU20051`. 2017TOU20051.

Els Lefever, Marjan Van de Kauter, and Véronique Hoste. Evaluation of automatic hypernym extraction from technical corpora in english and dutch. In *9th International Conference on Language Resources and Evaluation (LREC)*, pages 490–497. European Language Resources Association (ELRA), 2014.

J. Lehmann and Johanna Volker. *Perspectives on ontology learning*. IOS Press, 2014.

Roger Leitzke Granada. *Evaluation of methods for taxonomic relation extraction from text*. PhD thesis, Pontifícia Universidade Católica do Rio Grande do Sul and université de Toulouse, 2015.

Alessandro Lenci. Distributional semantics in linguistic and cognitive research. *From context to meaning : Distributional models of the lexicon in linguistics and cognitive science, special issue of the Italian Journal of Linguistics*, 20(1) :1–31, 2008.

Alessandro Lenci and Giulia Benotto. Identifying hypernyms in distributional semantic spaces. In **SEM 2012 : The First Joint Conference on Lexical and Computational Semantics – Volume 1 : Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 75–79, Montréal, Canada, 7-8 June 2012a. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/S12-1012`.

Alessandro Lenci and Giulia Benotto. Identifying hypernyms in distributional semantic spaces. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1 : Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 75–79. Association for Computational Linguistics, 2012b.

A. Maedche and S. Staab. Discovering conceptual relations from text. In W. Horn, editor, *ECAI 2000. Proceedings of the 14th European Conference on Artificial Intelligence*, pages 321–325. Amsterdam : IOS Press, 2000.

Véronique Malaisé, Pierre Zweigenbaum, and Bruno Bachimont. Detecting semantic relations between terms in definitions. In Sophia Ananadiou and Pierre Zweigenbaum, editors, *COLING 2004 CompuTerm 2004 : 3rd International Workshop on Computational Terminology*, pages 55–62, Geneva, Switzerland, August 29 2004. COLING. URL `http://www.aclweb.org/anthology/W04-1807`.

Jose L Martinez-Rodriguez, Aidan Hogan, and Ivan Lopez-Arevalo. Information extraction meets the semantic web : A survey. *Semantic Web*, online (Preprint) :1–81, 2018. URL `http://www.semantic-web-journal.net/content/information-extraction-meets-semantic-web-survey-0`.

Olena Medelyan, David N. Milne, Catherine Legg, and Ian H. Witten. Mining meaning from wikipedia. *Int. J. Hum.-Comput. Stud.*, 67(9) :716–754, 2009. doi : 10.1016/j.ijhcs.2009.05.004. URL `https://doi.org/10.1016/j.ijhcs.2009.05.004`.

Ingrid Meyer. Extracting knowledge-rich contexts for terminography - a conceptual and methodological framework. In Didier Bourigault, Christian Jacquemin, and Marie-Claude L'Homme, editors, *Recent Advances in Computational Terminology*, pages 279—302. IosPress, 2001.

Ingrid Meyer, Douglas Skuce, Lynne Bowker, and Karen Eck. Towards a new generation of terminological resources : An experiment in building a terminological knowledge base. In *13th International Conference on Computational Linguistics (COLING'92)*, pages 956–960, 1992.

Bonan Min, Ralph Grishman, Li Wan, Chang Wang, and David Gondek. Distant supervision for relation extraction with an incomplete knowledge base. In *HLT-NAACL*, pages 777–782, 2013.

Marvin Minsky. A framework for representing knowledge. Technical report, Massachusetts Institute of Technology, Cambridge, MA, USA, 1974.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP : Volume 2-Volume 2*, pages 1003–1011, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-46-6. URL `http://dl.acm.org/citation.cfm?id=1690219.1690287`.

Tom M et al. Mitchell. Machine learning. 1997. *Burr Ridge, IL : McGraw Hill*, 45(37) :870–877, 1997.

Emmanuel Morin. Des patrons lexico-syntaxiques pour aider au dépouillement terminologiques. *Traitement Automatique des Langues*, 40(1) :143–166, 1999.

Emmanuel Morin and Christian Jacquemin. Automatic acquisition and expansion of hypernym links. *Computers and the Humanities*, 38(4) :363–396, 2004.

Mohamed Morsey, Jens Lehmann, Sören Auer, Claus Stadler, and Sebastian Hellmann. Dbpedia and the live extraction of structured data from wikipedia. *Program*, 46(2) :157–181, 2012.

Roberto Navigli and Simone Paolo Ponzetto. Babelnet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193 : 217–250, 2012.

Roberto Navigli and Paola Velardi. Learning word-class lattices for definition and hypernym extraction. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 1318–1327. Association for Computational Linguistics, 2010.

Claire Nédellec, Adeline Nazarenko, and Robert Bossy. Information extraction. In Steffen Staab and Rudi Studer, editors, *Handbook on Ontologies*, pages 663–685. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. ISBN 978-3-540-92673-3. doi : 10.1007/978-3-540-92673-3_30. URL `https://doi.org/10.1007/978-3-540-92673-3_30`.

Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. A survey on open information extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3866–3878. Association for Computational Linguistics, 2018.

Sachin Pawar, Girish K Palshikar, and Pushpak Bhattacharyya. Relation extraction : A survey. *arXiv preprint arXiv :1712.05191*, 2017.

Georgios Petasis, Vangelis Karkaletsis, Georgios Paliouras, Anastasia Krithara, and Elias Zavitsanos. Ontology population and enrichment : State of the art. In Georgios Paliouras, Constantine D. Spyropoulos, and George Tsatsaronis, editors, *Knowledge-Driven Multimedia Information Extraction and Ontology Evolution - Bridging the Semantic Gap*, volume 6050 of *Lecture Notes in Computer Science*, pages 134–166. Springer, 2011. ISBN 978-3-642-20794-5. doi : 10.1007/978-3-642-20795-2_6. URL `https://doi.org/10.1007/978-3-642-20795-2\_6`.

Thierry Poibeau. *Extraction automatique d'information : Du texte brut au web sémantique*. Lavoisier, 2003. URL `https://hal.archives-ouvertes.fr/hal-00005506`. ISBN 2-7462-0610-2.

Christopher Ré. Software 2.0 and snorkel : Beyond hand-labeled data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 2876–2876, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5552-0. doi : 10.1145/3219819.3219937. URL `http://doi.acm.org/10.1145/3219819.3219937`.

Thomas Rebele, Fabian Suchanek, Johannes Hoffart, Joanna Biega, Erdal Kuzey, and Gerhard Weikum. Yago : A multilingual knowledge base from wikipedia, wordnet, and geonames. In Paul Groth, Elena Simperl, Alasdair Gray, Marta Sabou, Markus Krötzsch, Freddy Lecue, Fabian Flöck, and Yolanda Gil, editors, *The Semantic Web – ISWC 2016*, pages 177–185, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46547-0.

Sebastian Riedel, Limin Yao, and Andrew McCallum. Modeling relations and their mentions without labeled text. In *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases : Part III*, ECML PKDD'10, pages 148–163, Berlin, Heidelberg, 2010. Springer-Verlag. ISBN 3-642-15938-9, 978-3-642-15938-1. URL `http://dl.acm.org/citation.cfm?id=1889788.1889799`.

Enrico Santus, Alessandro Lenci, Qin Lu, and Sabine Schulte im Walde. Chasing hypernyms in vector spaces with entropy. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2 : Short Papers*, pages 38–42, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/E14-4008`.

Gwendolijn Schropp, Els Lefever, and Véronique Hoste. A combined pattern-based and distributional approach for automatic hypernym detection in dutch. In Galia Angelova, Kalina Bontcheva, and Ruslan Mitkov, editors, *RANLP*, pages 593–600. RANLP 2013 Organising Committee / ACL, 2013. URL `http://dblp.uni-trier.de/db/conf/ranlp/ranlp2013.html#SchroppLH13`.

Patrick Séguéla and Nathalie Aussenac-Gilles. Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine. In *Conférence ingénierie des connaissances*, pages 79–88, 1999.

Stuart Shapiro. A net structure for semantic information storage, deduction and retrieval. In *Proc. IJCAI-71*, pages 512–523, 09 1971.

Anuj Sharma, Vassilis Virvilis, Tina Lekka, and Christos Andronis. Binary relation extraction from biomedical literature using dependency trees and svms. *bioRxiv*, 2016.

Amit Singhal. Introducing the knowledge graph : Things, not strings, 2012-05-16. URL `https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html`.

Douglas Skuce and Ingrid Meyer. Concept analysis and terminology : A knowledge-based approach to documentation. In *Proceedings of the Thirteenth International Conference on Computational Linguistics (COLING 90)*, pages 56–58, 1990.

Alisa Smirnova and Philippe Cudré-Mauroux. Relation extraction using distant supervision : A survey. *ACM Comput. Surv.*, 51(5) :1–35, 2018.

J. F. Sowa. *Conceptual Structures : Information Processing in Mind and Machine*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1984. ISBN 0-201-14472-7.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago : A core of semantic knowledge unifying wordnet and wikipedia. In *Proceedings of the 16th International Conference on World Wide Web*, WWW'07, pages 697–706, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-654-7. doi : 10.1145/1242572.1242667. URL `http://doi.acm.org/10.1145/1242572.1242667`.

Asuka Sumida and Kentaro Torisawa. Hacking wikipedia for hyponymy relation acquisition. In *IJCNLP*, volume 8, pages 883–888. Citeseer, 2008.

Patrick Séguéla. *Construction de modèles de connaissances par analyse lingustiques de relations lexicales dans les documents techniques*. PhD thesis, Thèse de doctorat Ecole doctorale d'informatique et mathématiques de l'Université Toulouse 3, 2001. URL `http://www.theses.fr/2001TOU30210`. dirigée par Borillo Mario 2001TOU30210.

Peter D Turney, Patrick Pantel, et al. From frequency to meaning : Vector space models of semantics. *Journal of artificial intelligence research*, 37(1) :141–188, 2010.

Chengyu Wang, Xiaofeng He, and Aoying Zhou. A short survey on taxonomy learning from text corpora : Issues, resources and recent advances. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1190–1203, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi : 10.18653/v1/D17-1123. URL `https://www.aclweb.org/anthology/D17-1123`.

Peng Xu and Denilson Barbosa. Connecting language and knowledge with heterogeneous representations for neural relation extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3201–3206, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/N19-1323`.

Shubin Zhao and Ralph Grishman. Extracting relations with integrated information using kernel methods. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 419–426. Association for Computational Linguistics, 2005.

GuoDong Zhou, Jian Su, Jie Zhang, and Min Zhang. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 427–434, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. doi : 10.3115/1219840.1219893. URL `https://doi.org/10.3115/1219840.1219893`.